When and How Does In-Distribution Label Help Out-of-Distribution Detection?

Xuefeng Du [1]   Yiyou Sun [1]   Yixuan Li [1]

Abstract

Detecting data points deviating from the training distribution is pivotal for ensuring reliable machine learning. Extensive research has been dedicated to the challenge, spanning classical anomaly detection techniques to contemporary out-of-distribution (OOD) detection approaches. While OOD detection commonly relies on supervised learning from a labeled in-distribution (ID) dataset, anomaly detection may treat the entire ID data as a single class and disregard ID labels. This fundamental distinction raises a significant question that has yet to be rigorously explored: *when and how does ID label help OOD detection?* This paper bridges this gap by offering a formal understanding to theoretically delineate the impact of ID labels on OOD detection. We employ a graph-theoretic approach, rigorously analyzing the separability of ID data from OOD data in a closed-form manner. Key to our approach is the characterization of data representations through spectral decomposition on the graph. Leveraging these representations, we establish a provable error bound that compares the OOD detection performance with and without ID labels, unveiling conditions for achieving enhanced OOD detection. Lastly, we present empirical results on both simulated and real datasets, validating theoretical guarantees and reinforcing our insights. Code is publicly available at https://github.com/deeplearning-wisc/id_label.

1. Introduction

When deployed in the real world, machine learning models often encounter unfamiliar data points that fall outside the distribution of the observed data. This problem has been studied extensively, dating from the classical anomaly detec-

tion methods (Chandola et al., 2009; Ahmed & Courville, 2020; Han et al., 2022) to contemporary out-of-distribution (OOD) approaches (Liu et al., 2020b; Yang et al., 2021b; Fang et al., 2022).

While both anomaly detection and OOD detection share the goal of identifying test-time input that deviates from the training distribution, a crucial distinction lies in the *usage of in-distribution (ID) labels in training time*. Specifically, classical anomaly detection may disregard ID labels (Yang et al., 2021b), treating the entire ID dataset as a single class. In contrast, OOD detection commonly relies on supervised learning from a labeled ID dataset. It is reasonable to hypothesize that incorporating ID labels during training might influence the resulting feature representations, potentially leading to distinct capabilities in separating ID from OOD samples during test time. This raises a significant question that has yet to be rigorously explored in the field:

RQ: When and how does ID label help OOD detection?

Answering this question offers the fundamental key to understanding and bridging two highly related fields of anomaly detection and OOD detection. In pursuit of this objective, we provide a formal understanding to theoretically delineate the influence of ID labels on OOD detection. We base our analysis on a graph-theoretic approach by modeling the ID data via a graph, where the vertices are all the data points and edges encode the similarity among data. This analytical framework is well-suited for our investigation, as data points' representation similarity can differ between the self-supervised and supervised learning setting, contingent upon the availability of ID labels. For instance, when ID labels are present, the supervisory signal facilitates connecting points belonging to the same class, resulting in each class manifesting as a distinct connected sub-graph. In both cases (with or without ID labels), the sub-structures can be revealed by performing spectral decomposition on the graph and can be expressed equivalently as a contrastive learning objective on neural net representations (expounded further in Section 3). Importantly, these learned feature representations allow us to rigorously analyze the separability of ID data from OOD data in a closed-form manner.

Based on the analytical framework, we provide a formal error bound in Theorem 1, comparing the OOD detection performance with and without the inclusion of ID labels.
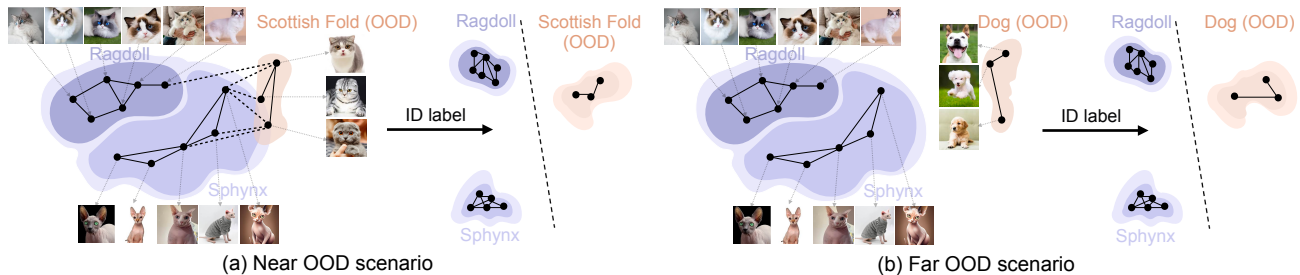
[1]Department of Computer Sciences, UW-Madison. Correspondence to: Yixuan Li <sharonli@cs.wisc.edu>.

Figure 1: Intuitive example on the ID labels' impact on OOD detection. (a) In the near OOD scenario where the OOD data connects densely with the ID data, without ID labels, the neural network produces indistinguishable embeddings for the ID (Ragdoll and Sphynx class) and OOD data (Scottish Fold class). By harnessing the power of the ID labeling information, the model learns more distinguishable embeddings that help ID vs. OOD separation. (b) In the far OOD scenario (Dog class), ID labels can be less beneficial because the representations learned in an unsupervised manner can already be separable between ID vs OOD.

This theorem reveals sufficient conditions for achieving improved OOD detection performance by leveraging ID labels. To establish the error bound, we first calculate the closed-form solution of the ID and OOD representations based on the graph factorization and then quantify the OOD detection performance by linear probing error. As a result, Theorem 1 demonstrates that the difference in the OOD detection performance with and without ID labels can be lower bounded by a function of the adjacency matrix of ID data as well as the OOD-ID connectivity. Furthermore, we offer intuitive interpretations in Theorem 2, and show that the ID labels are the most beneficial when either: (i) the OOD data is relatively close to the ID data, which is also known as the near OOD scenario; (ii) the ID data are connected sparsely without ID labels; and (iii) the semantic connection between each ID data to the labeled ID data from different ID classes is sufficiently large. To help readers understand the key insights of our theory, we provide a simple intuitive example in Figure 1, which demonstrates how adding ID labels in the near OOD scenario can lead to a greater benefit compared to the far OOD scenario.

Lastly, we provide empirical verifications to support our theory. In particular, we compare the OOD detection performance with and without ID labels using both simulated data and real-world datasets (Section 5). The result aligns with our theoretical guarantee, showcasing the benefits of the ID label information under proper conditions. For example, the OOD detection result can be improved by 12.3% (AUROC) in the near OOD scenario compared to 6.06% in the far OOD scenario on CIFAR100, validating our theory.

Our main contributions are summarized as follows:

- We study an important but underexplored problem: when and how in-distribution labels can help OOD detection. Our exposition has fundamental value in understanding and bridging the two highly related fields of anomaly detection and OOD detection.

- We provide an analytical framework based on graph formulation to characterize the ID and OOD representations. Based on that, we analyze the error bound for ID vs. OOD separation with and without ID labels and investigate the necessary conditions for which the labeling information can bring the most benefits.

- We present empirical analysis on both simulated and real-world datasets to verify and support our theory. The observation in practice echoes and reinforces our theoretical insights.

2. Problem Setup

Let $\mathcal{X}$ be the input space, and $\mathcal{Y} = \{1, ..., c\}$ be the label space for ID data. Given an unknown ID joint distribution $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$ defined over $\mathcal{X} \times \mathcal{Y}$, the *labeled ID data* $\mathcal{S}_{\text{id}}^{(l)} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ are drawn independently and identically from $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$. Alternatively, the *unlabeled ID data* $\mathcal{S}_{\text{id}}^{(u)} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ is drawn from $\mathbb{P}_{\text{id}}$, which is the marginal distribution of $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$ on $\mathcal{X}$. Furthermore, we denote $\mathbb{P}_{l_i}$ the distribution of labeled data with label $i \in \mathcal{Y}$.

Definition 1 (Out-of-Distribution Detection w/ ID Labels). *Given labeled ID data $\mathcal{S}_{\text{id}}^{(l)}$, the aim is to learn a predictor $\mathbf{g} : \mathcal{X} \to \mathcal{Y} \cup \{\text{ood}\}$ such that for any test data $\mathbf{x}$: 1) if $\mathbf{x}$ is drawn from $\mathbb{P}_{\text{id}}$, then the model classifies $\mathbf{x}$ into one of ID classes $\mathcal{Y}$, and 2) if $\mathbf{x}$ is drawn from another distribution $\mathbb{P}_{\text{ood}}$ with unknown OOD class, then $\mathbf{g}$ can detect $\mathbf{x}$ as OOD data (Fang et al., 2022).*

Definition 2 (Out-of-Distribution Detection w/o ID Labels). *The definition is similar to above, except that we are using unlabeled ID data $\mathcal{S}_{\text{id}}^{(u)} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ to learn the binary predictor $\mathbf{g}$. This is in accordance with the classical anomaly detection problem (Chandola et al., 2009).*

3. Analysis Framework

Overview of rationale. In this section, we introduce our analytical framework, which allows us to formalize and understand the OOD detection performance in two cases: **(1)** learning without ID labels, and **(2)** learning with ID labels, respectively. Our analytical framework models the ID data via a graph, where the vertices are all the data points, and edges encode the similarity among data (Section 3.1). The similarity can be defined in either a self-supervised or supervised manner, contingent on the availability of the ID labels. For example, when ID labels are present, the supervision signal can help connect points belonging to the same class, so that each class emerges clearly as a connected sub-graph. In both cases, the sub-structures can be revealed by performing spectral decomposition on the graph and can be expressed equivalently as a contrastive learning objective on neural net representations (Section 3.2). Importantly, these learned feature representations allow us to rigorously analyze the separability of ID data from OOD data in a closed form. Since the features learned can be directly impacted by the presence or absence of ID labels, the OOD detection performance can vary accordingly.

3.1. Graph Formulation

We start by formally defining the graph and adjacency matrix. For notation clarity, we use $\bar{\mathbf{x}}$ to indicate the natural sample (raw inputs without augmentation). Given an $\bar{\mathbf{x}}$, we use $\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}})$ to denote the probability of $\mathbf{x}$ being augmented from $\bar{\mathbf{x}}$. For instance, when $\bar{\mathbf{x}}$ represents an image, $\mathcal{T}(\cdot|\bar{\mathbf{x}})$ can be the distribution of common augmentations (Chen et al., 2020) such as Gaussian blur, color distortion, and random cropping. The augmentation allows us to define a general population space $\mathcal{X}_{\text{id}}$, which contains all the original ID data points along with their augmentations. We denote the cardinality of the population space with $|\mathcal{X}_{\text{id}}| = N$.

We define the graph $G(\mathcal{X}_{\text{id}}, \zeta)$ over the finite vertex set $\mathcal{X}_{\text{id}}$ with edge weights ζ. To define edge weights ζ, we consider two cases: (1) self-supervised connectivity $\zeta^{(u)}$ by treating all points in $\mathcal{X}_{\text{id}}$ as entirely unlabeled, and (2) supervised connectivity $\zeta^{(l)}$ by utilizing labeling information of ID data.

Definition 3 (Unlabeled case (u)). *When all ID points are unlabeled, two samples ($\mathbf{x}$, $\mathbf{x}^+$) are considered a **positive pair** if $\mathbf{x}$ and $\mathbf{x}^+$ are augmented from the same image $\bar{\mathbf{x}} \sim \mathbb{P}_{\text{id}}$. For any two augmented data $\mathbf{x}, \mathbf{x}' \in \mathcal{X}_{\text{id}}$, the edge weight $\zeta^{(u)}_{\mathbf{x}\mathbf{x}'}$ is defined as the marginal probability of generating the pair (HaoChen et al., 2021):*

$$\zeta^{(u)}_{\mathbf{x}\mathbf{x}'} \triangleq \mathbb{E}_{\bar{\mathbf{x}} \sim \mathbb{P}_{\text{id}}} \mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}) \mathcal{T}(\mathbf{x}'|\bar{\mathbf{x}}). \tag{1}$$

The magnitude of $\zeta^{(u)}_{\mathbf{x}\mathbf{x}'}$ indicates the "positiveness" or similarity between $\mathbf{x}$ and $\mathbf{x}'$.

Alternatively, when having access to the labeling information for ID data, we can define the edge weight by adding additional supervised connectivity to the graph.

Definition 4 (**Labeled case** (l)). *When all ID points are labeled, two samples ($\mathbf{x}$, $\mathbf{x}^+$) are considered a **positive pair** if $\mathbf{x}$ and $\mathbf{x}^+$ are augmented from two labeled samples $\bar{\mathbf{x}}_l$ and $\bar{\mathbf{x}}'_l$ with the same ID class i. The overall edge weight for any pair of data $(\mathbf{x}, \mathbf{x}')$ is given by:*

$$\zeta^{(l)}_{\mathbf{x}\mathbf{x}'} = \phi_u \cdot \zeta^{(u)}_{\mathbf{x}\mathbf{x}'} + \phi_l \cdot \sum_{i \in \mathcal{Y}} \mathbb{E}_{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}} \mathbb{E}_{\bar{\mathbf{x}}'_l \sim \mathbb{P}_{l_i}} \mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l) \mathcal{T}(\mathbf{x}'|\bar{\mathbf{x}}'_l),$$

where ϕ_u, ϕ_l are the weight coefficients. Compared to the unlabeled case, the second term strengthens the connectivity for points belonging to the same class.

Definition 5 (Adjacency matrix for unlabeled ID data). *We define the adjacency matrix $\mathbf{A}^{(u)}$ with entry value $\zeta^{(u)}_{\mathbf{x}\mathbf{x}'}$ for each $(\mathbf{x}, \mathbf{x}')$ pair. Further, $\zeta^{(u)}_{\mathbf{x}} = \sum_{\mathbf{x}' \in \mathcal{X}} \zeta^{(u)}_{\mathbf{x}\mathbf{x}'}$ denotes the total edge weights connected to a vertex $\mathbf{x}$.*

Definition 6 (Adjacency matrix for labeled ID data). *Similarly, we define the adjacency matrix for labeled ID data $\mathbf{A}^{(l)}$ with entry value $\zeta^{(l)}_{\mathbf{x}\mathbf{x}'}$ for each $(\mathbf{x}, \mathbf{x}')$ pair and $\zeta^{(l)}_{\mathbf{x}} = \sum_{\mathbf{x}' \in \mathcal{X}} \zeta^{(l)}_{\mathbf{x}\mathbf{x}'}$.*

As a standard technique in graph theory (Chung, 1997), we use the *normalized adjacency matrix* of $G(\mathcal{X}_{\text{id}}, \zeta)$:

$$\tilde{\mathbf{A}} \triangleq \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}, \tag{2}$$

where $\mathbf{A}$ can be instantiated by either $\mathbf{A}^{(u)}$ or $\mathbf{A}^{(l)}$ defined above. $\mathbf{D} \in \mathbb{R}^{N \times N}$ is the corresponding diagonal matrix with $\mathbf{D}_{\mathbf{x}\mathbf{x}} = \zeta^{(u)}_{\mathbf{x}}$ for unlabeled case and $\mathbf{D}_{\mathbf{x}\mathbf{x}} = \zeta^{(l)}_{\mathbf{x}}$ for labeled case. The normalization balances the degree of each node, reducing the influence of vertices with very large degrees. The normalized adjacency matrix allows us to perform spectral decomposition as we show next.

3.2. Learning Representations Based on Graph Spectral

In this section, we perform spectral decomposition or spectral clustering (Ng et al., 2001)—a classical approach to graph partitioning—to the adjacency matrices defined above. This process forms a matrix where the top-k eigenvectors are the columns and *each row of the matrix can be viewed as a k-dimensional representation of an example*. The resulting feature representations enable us to rigorously analyze the separability of ID data from OOD data in a closed form, and formally compare the OOD detection error under two scenarios either with and without ID labels (in Section 4).

Specifically, taking the labeled case as an example, we consider the following optimization, which performs low-rank matrix approximation on the adjacency matrix $\tilde{\mathbf{A}}^{(l)}$:

$$\min_{\mathbf{F}^{(l)} \in \mathbb{R}^{N \times k}} \mathcal{L}(\mathbf{F}^{(l)}, \mathbf{A}^{(l)}) \triangleq \left\| \tilde{\mathbf{A}}^{(l)} - \mathbf{F}^{(l)} \mathbf{F}^{(l)\top} \right\|_F^2, \tag{3}$$

where $\|\cdot\|_F$ denotes the matrix Frobenious norm. According to the Eckart–Young–Mirsky theorem (Eckart & Young, 1936), the minimizer of this loss function is $\mathbf{F}_k^{(l)} \in \mathbb{R}^{N \times k}$ such that $\mathbf{F}_k^{(l)}\mathbf{F}_k^{(l)\top}$ contains the top-k components of $\tilde{\mathbf{A}}^{(l)}$'s eigen decomposition.

A surrogate objective. In practice, directly solving objective 3 can be computationally expensive for an extremely large matrix. To circumvent this, the feature representations can be equivalently recovered by minimizing the following contrastive learning objective (Sun et al., 2023a;b) as shown in Lemma 1, which can be efficiently trained end-to-end using a neural net parameterized by $\mathbf{w}$:

$$\mathcal{L}_{\text{labeled}}(\mathbf{h_w}) \triangleq -2\phi_l\mathcal{L}_1(\mathbf{h_w}) - 2\phi_u\mathcal{L}_2(\mathbf{h_w}) + \phi_l^2\mathcal{L}_3(\mathbf{h_w}) + 2\phi_l\phi_u\mathcal{L}_4(\mathbf{h_w}) + \phi_u^2\mathcal{L}_5(\mathbf{h_w}), \quad (4)$$

where $\mathbf{h_w} : \mathcal{X}_{\text{id}} \mapsto \mathbb{R}^k$ denotes the feature representation,

$$\mathcal{L}_1(\mathbf{h_w}) = \sum_{i \in \mathcal{Y}} \underset{\substack{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}, \bar{\mathbf{x}}_l' \sim \mathbb{P}_{l_i}, \\ \mathbf{x} \sim \mathcal{T}(\cdot|\bar{\mathbf{x}}_l), \mathbf{x}^+ \sim \mathcal{T}(\cdot|\bar{\mathbf{x}}_l')}}{\mathbb{E}} \left[\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}^+\right)\right],$$

$$\mathcal{L}_2(\mathbf{h_w}) = \underset{\substack{\bar{\mathbf{x}}_u \sim \mathbb{P}_{\text{id}}, \\ \mathbf{x} \sim \mathcal{T}(\cdot|\bar{\mathbf{x}}_u), \mathbf{x}^+ \sim \mathcal{T}(\cdot|\bar{\mathbf{x}}_u)}}{\mathbb{E}} \left[\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}^+\right)\right],$$

$$\mathcal{L}_3(\mathbf{h_w}) = \sum_{i,j \in \mathcal{Y}} \underset{\substack{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}, \bar{\mathbf{x}}_l' \sim \mathbb{P}_{l_j}, \\ \mathbf{x} \sim \mathcal{T}(\cdot|\bar{\mathbf{x}}_l), \mathbf{x}^- \sim \mathcal{T}(\cdot|\bar{\mathbf{x}}_l')}}{\mathbb{E}} \left[\left(\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\right],$$

$$\mathcal{L}_4(\mathbf{h_w}) = \sum_{i \in \mathcal{Y}} \underset{\substack{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}, \bar{\mathbf{x}}_u \sim \mathbb{P}_{\text{id}}, \\ \mathbf{x} \sim \mathcal{T}(\cdot|\bar{\mathbf{x}}_l), \mathbf{x}^- \sim \mathcal{T}(\cdot|\bar{\mathbf{x}}_u)}}{\mathbb{E}} \left[\left(\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\right],$$

$$\mathcal{L}_5(\mathbf{h_w}) = \underset{\substack{\bar{\mathbf{x}}_u \sim \mathbb{P}_{\text{id}}, \bar{\mathbf{x}}_u' \sim \mathbb{P}_{\text{id}}, \\ \mathbf{x} \sim \mathcal{T}(\cdot|\bar{\mathbf{x}}_u), \mathbf{x}^- \sim \mathcal{T}(\cdot|\bar{\mathbf{x}}_u')}}{\mathbb{E}} \left[\left(\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\right].$$

Interpretation. At a high level, $\mathcal{L}_1$, $\mathcal{L}_2$ push embeddings of *positive pairs* to be closer while $\mathcal{L}_3$, $\mathcal{L}_4$ and $\mathcal{L}_5$ pull away embeddings of *negative pairs*. Particularly, $\mathcal{L}_1$ samples two random augmentation views of two images from labeled data with the *same* label. $\mathcal{L}_2$ samples two views from the same image in $\mathcal{X}_{\text{id}}$. For negative pairs, $\mathcal{L}_3$ uses two augmentation views from two labeled samples in $\mathcal{X}_l$ with *any* label. $\mathcal{L}_4$ uses two views of one sample in $\mathcal{X}_l$ and another one in $\mathcal{X}_{\text{id}}$. $\mathcal{L}_5$ uses two views from two random samples in $\mathcal{X}_{\text{id}}$.

Importantly, the contrastive loss allows drawing a theoretical equivalence between learned representations and the top-k singular vectors of $\tilde{\mathbf{A}}^{(l)}$, and facilitates theoretical understanding of the OOD detection on the data represented by $\tilde{\mathbf{A}}^{(l)}$. We formalize the equivalence below.

Lemma 1 (Theoretical equivalence between two objectives).
We define each row $\mathbf{f}_{\mathbf{x}}^\top$ of $\mathbf{F}^{(l)}$ as a scaled version of learned feature representation $\mathbf{h_w}$, with $\mathbf{f_x} = \sqrt{\zeta_{\mathbf{x}}}\mathbf{h_w}(\mathbf{x})$. Then minimizing the loss function $\mathcal{L}(\mathbf{F}^{(l)}, \mathbf{A}^{(l)})$ in Equation 3 is

equivalent to minimizing the surrogate loss in Equation 4. Full proof is in Appendix Section D.1.

Remark 1. We can extend the contrastive learning objective in Equation 4 to the *unlabeled case* by setting the coefficient ϕ_l to 0 and keeping the remaining parts:

$$\mathcal{L}_{\text{unlabeled}}(\mathbf{h_w}) \triangleq -2\phi_u\mathcal{L}_2(\mathbf{h_w}) + \phi_u^2\mathcal{L}_5(\mathbf{h_w}). \quad (5)$$

The loss has been employed in prior works on spectral contrastive learning (Sun et al., 2023a;b), which analyzed problems such as novel category discovery and open-world semi-supervised learning. However, our paper focuses on the problem of OOD detection, which has fundamentally different learning goals. Accordingly, we derive novel theoretical analyses uniquely tailored to our problem focus (i.e., the impact of the ID label information), which we present next.

4. Theoretical Results

Based on the analytical framework, we now provide theoretical insights to the core question: ***when and how does ID label information help OOD detection?*** To answer this question, we start by deriving the closed-form solution of the representations for both ID and OOD data (Section 4.1), and then quantify the OOD detection performance by measuring the linear probing error (Section 4.2). Finally, we provide a formal bound contrasting the OOD detection performance with and without ID labels (Section 4.3).

4.1. Representation for ID and OOD Data

ID representations. We first derive the ID representations for the labeled case, which can be similarly derived for the unlabeled case. Specifically, one can train the neural network $\mathbf{h_w} : \mathcal{X}_{\text{id}} \to \mathbb{R}^k$ using the surrogate objective in Equation 4. Minimizing the loss yields representation $\mathbf{Z}^{(l)} \in \mathbb{R}^{N \times k}$, where each row vector $\mathbf{z}_i = \mathbf{h_w}(\mathbf{x}_i)$. According to Lemma 1, the closed-form solution for the representations is equivalent to performing spectral decomposition of the adjacency matrix. Thus, we have $\mathbf{F}_k^{(l)} = [\mathbf{D}^{(l)}]^{\frac{1}{2}}\mathbf{Z}^{(l)}$, where $\mathbf{F}_k^{(l)}\mathbf{F}_k^{(l)\top}$ contains the top-k components of $\tilde{\mathbf{A}}^{(l)}$'s SVD decomposition. We further denote the top-k singular vectors of $\tilde{\mathbf{A}}^{(l)}$ as $\mathbf{V}_k^{(l)} \in \mathbb{R}^{N \times k}$, so we have $\mathbf{F}_k^{(l)} = \mathbf{V}_k^{(l)}[\mathbf{\Sigma}_k^{(l)}]^{\frac{1}{2}}$, where $\mathbf{\Sigma}_k^{(l)}$ is a diagonal matrix of the top-k singular values of $\tilde{\mathbf{A}}^{(l)}$. By equalizing the two forms of $\mathbf{F}_k^{(l)}$, the closed-formed solution of the learned feature space is given by

$$\mathbf{Z}^{(l)} = [\mathbf{D}^{(l)}]^{-\frac{1}{2}}\mathbf{V}_k^{(l)}[\mathbf{\Sigma}_k^{(l)}]^{\frac{1}{2}}. \quad (6)$$

OOD representations. In post hoc OOD detection, the learning algorithm can only observe ID data in $\mathcal{X}_{\text{id}}$ and the corresponding adjacency matrix. Hence, a key challenge in
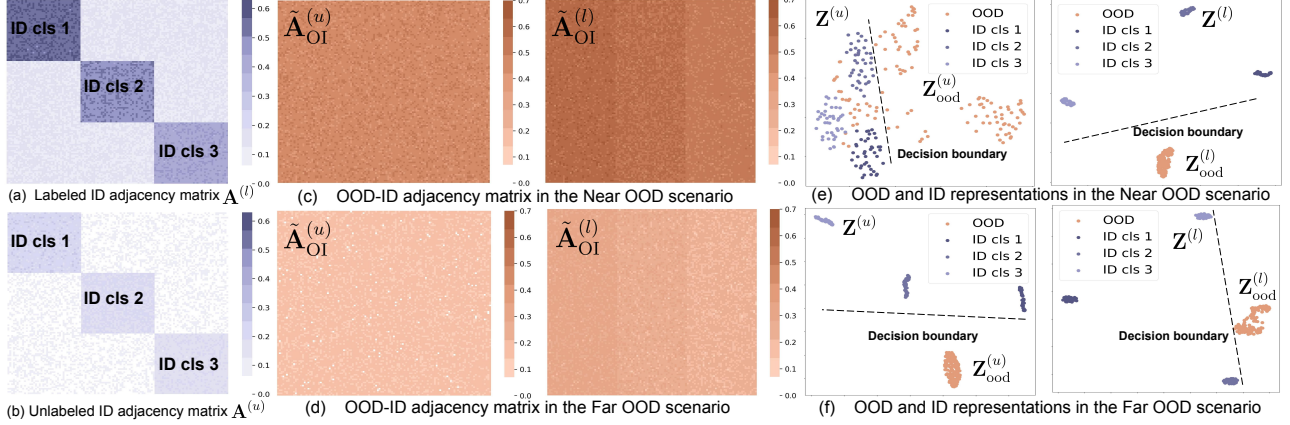
Figure 2: Example showcasing the contrast between adjacency matrices and representations w/ (l) and w/o (u) ID labels. (a) The ID adjacency matrix in the labeled case $\mathbf{A}^{(l)}$. (b) The ID adjacency matrix in the unlabeled case $\mathbf{A}^{(u)}$. Here darker color indicates denser connectivity. The contrast of the OOD-ID adjacency matrix $\tilde{\mathbf{A}}_{\mathrm{OI}}$ w/ and w/o ID labels in the near OOD and far OOD scenario is shown in (c) and (d), where the adjacency matrices have a larger Frobenius norm, i.e., $\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F = 60$ in the near OOD scenario and smaller norm in the far OOD scenario, i.e., $\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F = 24$. (e) Learned representations in the near OOD scenario, where the OOD representations are overlapped in the unlabeled case but become linearly separable from the ID representations in the labeled case. (d) Representations in the far OOD scenario. The ID and OOD representations can already be separable in the unlabeled case. The benefit of ID labels is marginal.

our framework is how to derive the OOD representations based on ID and OOD data connectivity in the input space. Unlike previous literature (Lee et al., 2018b), we refrain from making simplified assumptions in the feature representation space (although that makes analysis much easier). More realistically, we *characterize OOD data directly in the input space* by the adjacency matrix $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)} \in \mathbb{R}^{M \times N}$, where M is the number of OOD data points.

Each row in the matrix indicates the similarity between an OOD data w.r.t. all the ID samples. Depending on the characteristics of the OOD data, this matrix may be sparse if OOD data is far away from all the ID samples (e.g., far OOD), or can have dense entries if it is close to some ID classes (e.g., near OOD). Our characterization is thus general enough to enable analysis under different scenarios.

Now a question remains: how do we go from this matrix $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)}$ to a k-dimensional embedding for each OOD data? While a naive solution is to perform spectral decomposition on the stack of two matrices $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)}$ and $\tilde{\mathbf{A}}^{(l)}$, this violates the principle of post hoc OOD detection as it incurs retraining. Instead, we derive the embeddings of OOD vertices using existing ID embeddings $\mathbf{F}^{(l)}$ and the OOD-ID similarity. This can be achieved by solving the following optimization:

$$\min_{\mathbf{F}_{\mathrm{ood}}^{(l)} \in \mathbb{R}^{M \times k}} \left\| \tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)} - \mathbf{F}_{\mathrm{ood}}^{(l)} \mathbf{F}^{(l)\top} \right\|_F^2, \qquad (7)$$

where $\mathbf{F}_{\mathrm{ood}}^{(l)}$ denotes the OOD embeddings. Intuitively, the objective distills the similarity in the input space into the

representation space. For instance, it searches for an OOD representation, aligning it closely with ID representations when there is a dense connectivity between OOD and ID data in the adjacency matrix, and vice versa. Similar to the ID case, we have $\mathbf{Z}_{\mathrm{ood}}^{(l)} = [\mathbf{D}_{\mathrm{ood}}^{(l)}]^{-\frac{1}{2}} \mathbf{F}_{\mathrm{ood}}^{(l)}$, where $\mathbf{D}_{\mathrm{ood}}^{(l)}$ can be calculated in the same way as $\mathbf{D}^{(l)}$ based on $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)}$. Therefore, the analytic form of the OOD representations from the neural network can be derived as

$$\mathbf{Z}_{\mathrm{ood}}^{(l)} = [\mathbf{D}_{\mathrm{ood}}^{(l)}]^{-\frac{1}{2}} \tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)} \mathbf{V}_k^{(l)} [\mathbf{\Sigma}_k^{(l)}]^{-\frac{1}{2}}. \qquad (8)$$

Detailed proof and the design rationale are in Appendix D.2.

Representation in the unlabeled case. For the unlabeled case, we can get the representations for ID and OOD data by replacing the matrices $\mathbf{V}_k^{(l)}, \mathbf{\Sigma}_k^{(l)}, \mathbf{D}^{(l)}, \mathbf{D}_{\mathrm{ood}}^{(l)}$ and $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)}$ in Equations 6 and 8 with the unlabeled version. We will show how the labeled and unlabeled representations are rigorously related in Appendix C.1.

An illustrative example. To contrast the adjacency matrices and the corresponding representations for the **u**nlabeled and **l**abeled cases, we simulate an example in Figure 2. The simulation is constructed with simplicity in mind, to facilitate understanding. Evaluations on complex high-dimensional data will be provided in Section 5. In particular, we base our analysis on the ID adjacency matrix as depicted in Figure 2 (a) and (b), which consists of three ID classes and 40 data points for each class. In the labeled case, the ID adjacency matrix has a denser connectivity pattern, especially for data that belongs to the same ID label. In Figure 2

(c) and (d), we compare the OOD-ID adjacency matrices with and without ID labels in two scenarios, i.e., near OOD where there are dense connections in $\tilde{\mathbf{A}}_{\text{OI}}$ and far OOD where the connectivity in $\tilde{\mathbf{A}}_{\text{OI}}$ is sparse.

Based on the graph, we further visualize in Figures 2 (e) and (f) the 2D data representations ($k = 2$, calculated by Equations 6 and 8) for the near OOD and far OOD scenarios. We observe that having different adjacency matrices can lead to significantly different data representations. We will provide theory to rigorously analyze the OOD detection performance and contrast between the labeled and unlabeled cases (Section 4.3). Details of the illustrative example are included in Appendix F.

4.2. Evaluation Target

With the closed-form representations for both ID and OOD, we evaluate OOD detection by linear probing error. The strategy is commonly used in representation learning (Chen et al., 2020). Specifically, the weight of a linear classifier is denoted as $\boldsymbol{\theta} \in \mathbb{R}^{k \times 2}$. The class prediction (ID vs. OOD) is given by $\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}) = \operatorname{argmax}_{i \in \{+, -\}}(\mathbf{h}_{\mathbf{w}}(\mathbf{x})\boldsymbol{\theta})_i$. Denote the set of ID and OOD features as $\mathbf{Z}_{\text{all}} \in \mathbb{R}^{(N+M) \times k} = [\mathbf{Z}^\top, \mathbf{Z}_{\text{ood}}^\top]^\top$ (either labeled (u) or unlabeled (l)), the linear probing error $R(\mathbf{Z}_{\text{all}})$ is given by the least error of all possible linear classifiers:

$$R(\mathbf{Z}_{\text{all}}) \triangleq \min_{\boldsymbol{\theta} \in \mathbb{R}^{k \times 2}} \mathbb{E}_{\mathbf{z} \in \mathbf{Z}_{\text{all}}} \mathbb{1}[y(\mathbf{z}) \neq \mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})], \qquad (9)$$

where $y(\mathbf{z})$ denotes indicates the ground-truth class of feature $\mathbf{z}$ (ID or OOD). With the definition, we can bound the linear probing error $R(\mathbf{Z}_{\text{all}})$ by the residual of the regression error as shown in Lemma 2 with proof in Appendix D.3.

Lemma 2. *Denote* $\mathbf{y} \in \mathbb{R}^{(N+M) \times 2}$ *as a matrix where each row contains the one-hot label for features in* $\mathbf{Z}_{\text{all}}$*. We have:*

$$R(\mathbf{Z}_{\text{all}}) \leq \frac{2}{N + M} \operatorname{Tr}\left(\left(\mathbf{I} - \mathbf{Z}_{\text{all}}\mathbf{Z}_{\text{all}}^\dagger\right)\mathbf{y}\mathbf{y}^\top\right). \qquad (10)$$

Here $\operatorname{Tr}(\cdot)$ denotes the trace operator. $\mathbf{Z}_{\text{all}}^\dagger$ is the Moore-Penrose inverse of matrix $\mathbf{Z}_{\text{all}}$. We denote this upper bound as $\overline{R}(\mathbf{Z}_{\text{all}})$, which is more tractable to analyze and behaves similarly to $R(\mathbf{Z}_{\text{all}})$ as shown in Appendix D.3. Therefore our subsequent analysis revolves around it.

4.3. Error Bound on OOD Detection Performance

With the evaluation target defined above, we now present the formal error bound on OOD detection performance by contrasting the labeled and unlabeled cases. As an overview, Theorem 1 will present the lower bound of linear probing error difference between the unlabeled and label case, along with an intuitive version in Theorem 2. We specify several mild assumptions and necessary notations for our theorems

in Appendix A. Due to space limitation, we omit unimportant constants and simplify the statements of our theorems. We defer the **full formal** statements in Appendix B. All proofs can be found in Appendix C.

Error difference between unlabeled and labeled cases. Formally, we investigate the following linear probing error difference between the unlabeled and labeled case:

$$\mathcal{G} = \overline{R}(\mathbf{Z}_{\text{all}}^{(u)}) - \overline{R}(\mathbf{Z}_{\text{all}}^{(l)}), \qquad (11)$$

where a larger error difference indicates that labeled ID data benefits OOD detection more substantially, and vice versa. The lower bound on $\mathcal{G}$ is given by the following theorem.

Theorem 1 (Lower bound of the linear probing error difference w/ and w/o ID labels)**.** *(Informal.) Suppose we have adjacency matrices* $\tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}^{(l)} \in \mathbb{R}^{N \times N}$ *and* $\tilde{\mathbf{A}}_{\text{OI}}^{(u)}, \tilde{\mathbf{A}}_{\text{OI}}^{(l)} \in \mathbb{R}^{M \times N}$ *for both the labeled and unlabeled cases. Under mild conditions, given positive constants* ϕ_l, C*, the error difference* $\mathcal{G}$ *in Equation 11 is lower bounded by*

$$\mathcal{G} \geq \frac{C\phi_l}{N + M}\epsilon(\mathfrak{p}, \mathfrak{q}, \tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}_{\text{OI}}^{(u)}), \qquad (12)$$

where $\mathfrak{q} \in \mathbb{R}^{N \times c}$ *with each column defined as* $(\mathfrak{q}_i)_{\mathbf{x}} = \mathbb{E}_{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l), \mathbf{x} \in \mathcal{X}_{\text{id}}$*. Similarly,* $\mathfrak{p} \in \mathbb{R}^{M \times c}$ *is defined as* $(\mathfrak{p}_i)_{\mathbf{x}} = \mathbb{E}_{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l), \mathbf{x} \in \mathcal{X}_{\text{ood}}$*. Semantically, each entry in* $\mathfrak{q}$ *means the connection magnitude from* $\mathbf{x}$ *to all ID data while each entry in* $\mathfrak{p}$ *is the connection from* $\mathbf{x}$ *to OOD data. Furthermore,*

$$\epsilon(\mathfrak{p}, \mathfrak{q}, \tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}_{\text{OI}}^{(u)}) = 2\sum_{i=1}^{c} \operatorname{Tr}\left(\mathfrak{p}_i\mathfrak{q}_i^\top \cdot \tilde{\mathbf{A}}_{\text{OI}}^{(u)\top}\right) +$$

$$\left(1 - \|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F^2\|\tilde{\mathbf{A}}^{(u)}\|_F^2\right)\sum_{i=1}^{c}\|\mathfrak{q}_i\|_F^2 +$$

$$r^2\left(\|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F^2\|\tilde{\mathbf{A}}^{(u)}\|_F^2\frac{2(\tau - k)}{\tau - 1} - 2\right) \cdot \sum_{i=1}^{c}\|\mathfrak{q}_i\|_1,$$

where τ *is a constant that measures the k-th spectral gap of matrix* $\tilde{\mathbf{A}}^{(u)}$*, i.e.,* $\boldsymbol{\lambda}_k^{(u)} \geq \tau\boldsymbol{\lambda}_{k+1}^{(u)}$ *and* $\boldsymbol{\lambda}_k^{(u)}$ *is the k-th largest singular value of* $\tilde{\mathbf{A}}^{(u)}$*. r is the maximum l_2 norm of the ID representations, i.e.,* $r = \max_{\mathbf{z} \in \mathbf{Z}^{(u)}}\|\mathbf{z}\|_2$*.*

Theorem 1 is a general characterization of the error difference in the labeled and unlabeled cases. To gain a better insight, we introduce Theorem 2 which provides intuition interpretations.

OOD category	OOD dataset	ID labels	FPR95↓	AUROC↑	LP error↓	FPR95↓	AUROC↑	LP error↓
			$\mathbb{P}_{\text{ood}}^{\text{test}} = \mathbb{P}_{\text{ood}}^{\text{LP}}$			$\mathbb{P}_{\text{ood}}^{\text{test}} \neq \mathbb{P}_{\text{ood}}^{\text{LP}}$		
FAR OOD	SVHN	-	$0.09^{\pm0.02}$	$99.96^{\pm0.01}$	$\mathbf{0.02^{\pm0.00}}$	$75.62^{\pm4.74}$	$77.09^{\pm2.17}$	$0.52^{\pm0.21}$
		+	$\mathbf{0.07^{\pm0.00}}$	$\mathbf{99.97^{\pm0.03}}$	$\mathbf{0.02^{\pm0.00}}$	$\mathbf{68.73^{\pm5.03}}$	$\mathbf{82.97^{\pm1.87}}$	$\mathbf{0.48^{\pm0.28}}$
	TEXTURES	-	$0.37^{\pm0.20}$	$99.80^{\pm0.15}$	$0.02^{\pm0.00}$	$86.35^{\pm2.47}$	$70.94^{\pm3.77}$	$1.05^{\pm0.17}$
		+	$\mathbf{0.24^{\pm0.19}}$	$\mathbf{99.86^{\pm0.11}}$	$\mathbf{0.01^{\pm0.00}}$	$\mathbf{86.44^{\pm0.58}}$	$\mathbf{75.15^{\pm3.36}}$	$\mathbf{0.95^{\pm0.10}}$
	PLACES365	-	$0.68^{\pm0.10}$	$99.98^{\pm0.00}$	$0.02^{\pm0.01}$	$75.35^{\pm2.78}$	$79.85^{\pm1.85}$	$0.83^{\pm0.27}$
		+	$\mathbf{0.44^{\pm0.06}}$	$\mathbf{99.99^{\pm0.01}}$	$0.02^{\pm0.01}$	$\mathbf{66.60^{\pm1.61}}$	$\mathbf{85.03^{\pm3.19}}$	$\mathbf{0.74^{\pm0.09}}$
	LSUN-RESIZE	-	$0.24^{\pm0.03}$	$99.95^{\pm0.03}$	$0.02^{\pm0.00}$	$83.57^{\pm2.89}$	$77.57^{\pm5.21}$	$0.84^{\pm0.21}$
		+	$\mathbf{0.24^{\pm0.01}}$	$99.91^{\pm0.07}$	$0.02^{\pm0.01}$	$\mathbf{74.13^{\pm4.95}}$	$\mathbf{82.71^{\pm1.64}}$	$\mathbf{0.75^{\pm0.14}}$
	LSUN-C	-	$1.68^{\pm0.36}$	$99.20^{\pm0.17}$	$0.05^{\pm0.01}$	$63.42^{\pm6.16}$	$83.43^{\pm3.48}$	$0.82^{\pm0.38}$
		+	$\mathbf{1.04^{\pm0.41}}$	$\mathbf{99.35^{\pm0.08}}$	$\mathbf{0.04^{\pm0.02}}$	$\mathbf{51.36^{\pm2.26}}$	$\mathbf{89.49^{\pm1.91}}$	$\mathbf{0.72^{\pm0.41}}$
NEAR OOD	CIFAR10	-	$62.20^{\pm3.49}$	$85.93^{\pm1.72}$	$0.27^{\pm0.11}$	$94.54^{\pm0.79}$	$55.42^{\pm1.64}$	$0.97^{\pm0.18}$
		+	$\mathbf{58.28^{\pm2.90}}$	$\mathbf{89.01^{\pm0.98}}$	$\mathbf{0.19^{\pm0.05}}$	$\mathbf{91.07^{\pm3.28}}$	$\mathbf{67.72^{\pm0.97}}$	$\mathbf{0.85^{\pm0.32}}$

Table 1: OOD detection results w/ and w/o ID labels (CIFAR100 as ID). Mean and std are estimated on three different runs. Better results are highlighted in bold. "+,-" denotes the labeled and unlabeled case. "LP error" denotes the error of linear probing. ↑ indicates larger values are better, and ↓ indicates smaller values are better. The table shows that (1) the ID labeling information helps OOD detection, especially in the near OOD scenario and when $\mathbb{P}_{\text{ood}}^{\text{test}} = \mathbb{P}_{\text{ood}}^{\text{LP}}$; Moreover, our observations can generalize to the case where the OOD distribution in linear probing is not the same as that in actual testing ($\mathbb{P}_{\text{ood}}^{\text{test}} \neq \mathbb{P}_{\text{ood}}^{\text{LP}}$), showcasing the generality of our theory.

Theorem 2 (Intuitive version of Theorem 1). *Under the same conditions in Theorem 1, assume the k-th spectral gap of $\tilde{\mathbf{A}}^{(u)}$ is sufficiently large, i.e., $\tau > k$, then the main error component $\epsilon(\mathfrak{p}, \mathfrak{q}, \tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}_{\text{OI}}^{(u)})$ in Equation 12 satisfies*

$$\epsilon(\mathfrak{p}, \mathfrak{q}, \tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}_{\text{OI}}^{(u)}) \geq \left[1 + \|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F^2 \left(2N^2 - \|\tilde{\mathbf{A}}^{(u)}\|_F^2\right)\right] \|\mathfrak{q}\|_F^2.$$

Interpretation and key insights. Theorem 2 shows that the optimal scenarios for achieving the greatest reduction in linear probing error—signifying the most significant benefit by incorporating the ID labels, are when

1. The ID data are connected sparsely in the unlabeled case (i.e., $\|\tilde{\mathbf{A}}^{(u)}\|_F^2 < 2N^2$), which always holds because $\|\tilde{\mathbf{A}}^{(u)}\|_\infty < N$;

2. The OOD data is closely connected to the ID data (near OOD, i.e., $\|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F^2$ is relatively large);

3. The semantic connection between each ID data to the labeled ID data from different ID classes is sufficiently large (i.e., $\|\mathfrak{q}\|_F^2$ is large).

Moreover, the simplified bound also enables us to interpret the relationship of each key component with the error reduction $\mathcal{G}$. For example, 1) *the bound will monotonically increase when the connection within ID data in the unlabeled case becomes sparser (i.e., $\|\tilde{\mathbf{A}}^{(u)}\|_F^2 \downarrow$); 2) Since $\|\tilde{\mathbf{A}}^{(u)}\|_F^2$ is smaller than $2N^2$, strengthening the semantic connection from each ID data to the labeled data from different ID classes (i.e., $\|\mathfrak{q}\|_F^2 \uparrow$) is always helpful. Intuitively,* a larger $\|\mathfrak{q}\|_F^2$ means one ID data point is more likely to be augmented from another ID data; 3) If the ID and OOD data are closer ($\|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F^2 \uparrow$), the value of the bound will increase given the same $\mathfrak{q}$ and $\tilde{\mathbf{A}}^{(u)}$.

Verification of bound on the illustrative example. Our theoretical guarantee aligns well with the empirical results. For example, in Figure 2 (e) where the OOD data is densely connected to the ID data (near OOD case), the ID labels can better shape the ID and OOD representations compared to the unlabeled case, rendering them linearly separable. As a result, the linear probing error is reduced from 0.09 in the unlabeled case to 0 in the labeled case. In contrast, when the OOD data is far from the ID data (Figure 2 (f)), the representations for ID and OOD are already well separated in the unlabeled case, and thus the benefit of ID labels is relatively marginal. As a verification, the linear probing error remains 0 both with and without ID labels. Therefore, these observations align with our key insight on the effect of OOD-ID connection $\tilde{\mathbf{A}}_{\text{OI}}^{(u)}$ on the error difference for linear probing. Moreover, we provide additional visualization results on changing the Frobenius norm of the ID adjacency matrix $\tilde{\mathbf{A}}^{(u)}$ and the semantic connection $\mathfrak{q}$ in Appendix G.

5. Experiments on Real Datasets

In this section, we verify our theoretical results using real-world OOD detection benchmarks.

Experimental setup. For ID datasets, we use CIFAR10 and CIFAR100 (Krizhevsky et al., 2009). We first train the neural network on the ID data for 200 epochs with a ResNet-18 (He et al., 2016), using objective $\mathcal{L}_{\text{unlabeled}}$ and $\mathcal{L}_{\text{labeled}}$ for the unlabeled and labeled case, respectively. The

penultimate layer embedding dimension $k = 512$. We then extract the embeddings for the ID and OOD data and perform linear probing (50 epochs). We explore two different scenarios depending on whether the OOD in linear probing ($\mathbb{P}_{\text{ood}}^{\text{LP}}$) is the same as the test OOD ($\mathbb{P}_{\text{ood}}^{\text{test}}$). In the first scenario ($\mathbb{P}_{\text{ood}}^{\text{test}} = \mathbb{P}_{\text{ood}}^{\text{LP}}$), we use 75% of the OOD dataset for linear probing and the remaining for testing. Specifically, for far OOD test datasets, we use a suite of natural image datasets including TEXTURES (Cimpoi et al., 2014), SVHN (Netzer et al., 2011), PLACES365 (Zhou et al., 2017), and LSUN (Yu et al., 2015). For near OOD, we evaluate on CIFAR10 when CIFAR100 is ID and vice versa. In the second scenario ($\mathbb{P}_{\text{ood}}^{\text{test}} \neq \mathbb{P}_{\text{ood}}^{\text{LP}}$), we use 300K RANDOM IMAGES (Hendrycks et al., 2019) for linear probing and the other OOD datasets for evaluation. More experiment details are in Appendix H.

Evaluation metrics. We report the following metrics: (1) the false positive rate (FPR95↓) of OOD samples when the true positive rate of ID samples is 95%, (2) the area under the receiver operating characteristic curve (AUROC↑), and (3) the linear probing error (LP error ↓).

Experiment results. The results are shown in Table 1, which demonstrate that: (1) the ID labeling information helps OOD detection, especially in the near OOD scenario and when $\mathbb{P}_{\text{ood}}^{\text{test}} = \mathbb{P}_{\text{ood}}^{\text{LP}}$. For example, the AUROC is improved by 3.08% on CIFAR10 compared to 0.02% on SVHN, echoing our theoretical insights; (2) Our observations can generalize to the case where the OOD distribution in linear probing is not the same as that in actual testing ($\mathbb{P}_{\text{ood}}^{\text{test}} \neq \mathbb{P}_{\text{ood}}^{\text{LP}}$), where the AUROC increases by 12.3% compared to the unsupervised counterpart on CIFAR10, showcasing the flexibility and generality of our theory. Additional results on CIFAR10 as the ID dataset and the evaluation using post-hoc OOD detection score are shown in Appendix I.

Verification of bound. We verify the error difference $\mathcal{G}$ and its relationship to the Frobenius norm of the adjacency matrices $\tilde{\mathbf{A}}^{(u)}$ and $\tilde{\mathbf{A}}_{\text{OI}}^{(u)}$. Firstly, to verify how the value of $\mathcal{G}$ will change given a larger Frobenius norm of $\tilde{\mathbf{A}}_{\text{OI}}^{(u)}$, we compare the linear probing error on SVHN and CIFAR10 in Table 2 with and without ID labels, where the error difference on CIFAR10 (near OOD, with larger $\|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F$) is consistently larger than that on SVHN (far OOD).

In addition, we verify the relationship of $\|\tilde{\mathbf{A}}^{(u)}\|_F$ and the error difference (CIFAR10 as OOD) in Table 3. Specifically, we calculate the norm of the ID adjacency matrix from

Epochs	40	80	120	160	200	240
$\|\tilde{\mathbf{A}}^{(u)}\|_F \downarrow$	20191	19549	18939	16073	15509	14810
$\mathcal{G} \uparrow$	0.01	0.03	0.06	0.06	0.08	0.09

Table 3: Verification with different $\|\tilde{\mathbf{A}}^{(u)}\|_F$ (CIFAR100 as ID).

different training epochs and observe that the difference in linear probing error tends to increase with decreasing $\|\tilde{\mathbf{A}}^{(u)}\|_F$, which aligns with Theorem 2. Additional results and details are included in Appendix E. We further analyze the tightness of our bound in Appendix J.

6. Related Work

OOD detection has attracted a surge of interest in recent years (Fort et al., 2021; Yang et al., 2021b; Fang et al., 2022; Zhu et al., 2022; Ming et al., 2022a;c; Yang et al., 2022; Wang et al., 2022b; Galil et al., 2023; Djurisic et al., 2023; Zheng et al., 2023; Wang et al., 2022a; 2023b; Narasimhan et al., 2023; Yang et al., 2023; Uppaal et al., 2023; Zhu et al., 2023b;a; Ming & Li, 2023; Zhang et al., 2023; Ghosal et al., 2024). One line of work performs OOD detection by devising scoring functions, including confidence-based methods (Bendale & Boult, 2016; Hendrycks & Gimpel, 2017; Liang et al., 2018), energy-based score (Liu et al., 2020b; Wang et al., 2021; Wu et al., 2023), distance-based approaches (Lee et al., 2018b; Tack et al., 2020; Ren et al., 2021; Sehwag et al., 2021a; Sun et al., 2022; Du et al., 2022a; Ming et al., 2023; Ren et al., 2023), gradient-based score (Huang et al., 2021), and Bayesian approaches (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Maddox et al., 2019; Malinin & Gales, 2019; Wen et al., 2020; Kristiadi et al., 2020). Another line of work addressed OOD detection by training-time regularization (Bevandić et al., 2018; Malinin & Gales, 2018; Geifman & El-Yaniv, 2019; Hein et al., 2019; Meinke & Hein, 2020; Jeong & Kim, 2020; Liu et al., 2020a; van Amersfoort et al., 2020; Yang et al., 2021a; Wei et al., 2022; Du et al., 2022b; 2023; Wang et al., 2023a). For example, the model is regularized to produce lower confidence (Lee et al., 2018a) or higher energy (Liu et al., 2020b; Du et al., 2022c) on a set of clean OOD data (Hendrycks et al., 2019; Ming et al., 2022b), wild data (Zhou et al., 2021; Katz-Samuels et al., 2022; He et al., 2023; Bai et al., 2023; Du et al., 2024) and synthetic outliers (Du et al., 2023; Tao et al., 2023; Park et al., 2023). Additionally, a similar topic in a different domain, i.e., anomaly detection, often restricts the ID (normality) to be with a single class and with a different definition of the outliers (Chandola et al., 2009; Han et al., 2022).

Most OOD detection methods rely on the supervision of ID labels, and there have been prior works, such as (Tack et al., 2020; Sehwag et al., 2021b; Sun et al., 2022) that empirically verified that training with the ID labels can achieve a

OOD dataset	SVHN	C10
	FAR OOD	NEAR OOD
$\|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F \uparrow$	2334	2739
$\mathcal{G} \uparrow$	0.00	**0.08**

Table 2: Verification with different $\|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F$ (CIFAR100 as ID).

much better OOD detection performance compared to the unsupervised version, but a provable analysis of their impact is critical yet missing in the field.

OOD detection theory. Recent studies have begun to focus on the theoretical understanding of OOD detection. Fang et al. (2022) studied the generalization of OOD detection by PAC learning and they found a necessary condition for the learnability of OOD detection. Morteza & Li (2022) derived a novel OOD score and provided a provable understanding of the OOD detection result using that score. Du et al. (2024) theoretically studied the impact of unlabeled data for OOD detection. In contrast, we formally analyze the impact of ID labels on OOD detection, which has not been studied in the past.

Spectral graph theory is a classical research problem (Von Luxburg, 2007; Chung, 1997; Cheeger, 2015; Kannan et al., 2004; Lee et al., 2014; McSherry, 2001), which aims to partition the graph by studying the eigenspace of the adjacency matrix. Recently, it has been applied in different applications in machine learning (Ng et al., 2001; Shi & Malik, 2000; Blum, 2001; Zhu et al., 2003; Argyriou et al., 2005; Shaham et al., 2018). HaoChen et al. (2021) derived the spectral contrastive learning from the factorization of the graph's adjacency matrix, and provably understand unsupervised domain adaptation (Shen et al., 2022; HaoChen et al., 2022). Sun et al. (2023a;b) expanded the spectral contrastive learning approach to novel class discovery and open-world semi-supervised learning. Our focus is on the OOD detection problem, which differs from prior literature.

7. Conclusion

In this paper, we propose a novel analytical framework that studies the impact of ID labels on OOD detection. Our framework takes a graph-theoretic approach by modeling the ID data via a graph, which allows us to characterize the feature representations by performing spectral decomposition on the graph that can be expressed equivalently as a contrastive learning objective on neural net representation. Leveraging these representations, we establish a provable error bound that compares the OOD detection performance with and without ID labels, which reveals sufficient conditions for achieving improved OOD detection performance. Empirical observations further support our theoretical conclusions, showcasing the benefits of ID labeling information under proper conditions. We hope our work will inspire future research on the theoretical understanding of OOD detection. One promising direction is to analyze the setting where there is access to OOD samples, which belongs to an important branch of work in OOD detection literature.

Acknowledgement

Impact Statement

Our paper aims to improve the reliability and safety of modern machine learning models. From the theoretical perspective, our analysis can facilitate and deepen the understanding of the effect of in-distribution labels for OOD detection. In Appendix E and Section 5 of the main paper, we properly verify the necessary conditions and the value of our bound using real-world datasets. Hence, we believe our theoretical framework has a broad utility and significance.

From the practical side, our study can lead to direct benefits and societal impacts by deploying OOD detection techniques, particularly when practitioners need to determine how to better leverage the in-distribution labels, such as in safety-critical applications i.e., autonomous driving and healthcare data analysis. Our study does not involve any human subjects or violation of legal compliance. We do not anticipate any potentially harmful consequences to our work. Through our study and releasing our code, we hope to raise stronger research and societal awareness on the safe handling of out-of-distribution data in real-world settings.

References

Ahmed, F. and Courville, A. Detecting semantic anomalies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3154–3162, 2020.

Argyriou, A., Herbster, M., and Pontil, M. Combining graph laplacians for semi–supervised learning. *Advances in Neural Information Processing Systems*, 18, 2005.

Bai, H., Canal, G., Du, X., Kwon, J., Nowak, R. D., and Li, Y. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, 2023.

Baksalary, J. K. and Puntanen, S. An inequality for the trace of matrix product. *IEEE transactions on automatic control*, 37(2):239–240, 1992.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(85):2399–2434, 2006.

Bendale, A. and Boult, T. E. Towards open set deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1563–1572, 2016.

Bengio, Y., Paiement, J.-f., Vincent, P., Delalleau, O., Roux, N., and Ouimet, M. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Advances in neural information processing systems*, 16, 2003.

Bevandić, P., Krešo, I., Oršić, M., and Šegvić, S. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018.

Blum, A. Learning form labeled and unlabeled data using graph mincuts. In *Proc. 18th International Conference on Machine Learning, 2001*, 2001.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Cheeger, J. A lower bound for the smallest eigenvalue of the laplacian. In *Problems in analysis*, pp. 195–200. Princeton University Press, 2015.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Chung, F. R. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Djurisic, A., Bozanic, N., Ashok, A., and Liu, R. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations*, 2023.

Du, X., Gozum, G., Ming, Y., and Li, Y. Siren: Shaping representations for detecting out-of-distribution objects. In *Advances in Neural Information Processing Systems*, 2022a.

Du, X., Wang, X., Gozum, G., and Li, Y. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.

Du, X., Wang, Z., Cai, M., and Li, Y. Vos: Learning what you don't know by virtual outlier synthesis. In *Proceedings of the International Conference on Learning Representations*, 2022c.

Du, X., Sun, Y., Zhu, X., and Li, Y. Dream the impossible: Outlier imagination with diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Du, X., Fang, Z., Diakonikolas, I., and Li, Y. How does unlabeled data provably help out-of-distribution detection? In *Proceedings of the International Conference on Learning Representations*, 2024.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Fang, Y., Loparo, K., and Feng, X. Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39(12):2489–2490, 1994.

Fang, Z., Li, Y., Lu, J., Dong, J., Han, B., and Liu, F. Is out-of-distribution detection learnable? In *Advances in Neural Information Processing Systems*, 2022.

Fort, S., Ren, J., and Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, pp. 1050–1059, 2016.

Galil, I., Dabbah, M., and El-Yaniv, R. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. In *International Conference on Learning Representations*, 2023.

Geifman, Y. and El-Yaniv, R. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning*, pp. 2151–2159, 2019.

Ghosal, S. S., Sun, Y., and Li, Y. How to overcome curse-of-dimensionality for ood detection? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Greenbaum, A., Li, R.-c., and Overton, M. L. First-order perturbation theory for eigenvalues and eigenvectors. *SIAM review*, 62(2):463–482, 2020.

Han, S., Hu, X., Huang, H., Jiang, M., and Zhao, Y. Ad-bench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 35:32142–32159, 2022.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

HaoChen, J. Z., Wei, C., Kumar, A., and Ma, T. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *Advances in Neural Information Processing Systems*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, R., Li, R., Han, Z., Yang, X., and Yin, Y. Topological structure learning for weakly-supervised out-of-distribution detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4858–4866, 2023.

Hein, M., Andriushchenko, M., and Bitterwolf, J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of the International Conference on Learning Representations*, 2017.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. In *Proceedings of the International Conference on Learning Representations*, 2019.

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, 2021.

Jansen, A., Sell, G., and Lyzinski, V. Scalable out-of-sample extension of graph embeddings using deep neural networks. *arXiv preprint arXiv:1508.04422*, 2015.

Jeong, T. and Kim, H. Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification. *Advances in Neural Information Processing Systems*, 33:3907–3916, 2020.

Kannan, R., Vempala, S., and Vetta, A. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515, 2004.

Katz-Samuels, J., Nakhleh, J., Nowak, R., and Li, Y. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, 2022.

Kristiadi, A., Hein, M., and Hennig, P. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pp. 5436–5446, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Lee, J. R., Gharan, S. O., and Trevisan, L. Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)*, 61(6):1–30, 2014.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *Proceedings of the International Conference on Learning Representations*, 2018a.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31, 2018b.

Levin, K., Roosta, F., Mahoney, M., and Priebe, C. Out-of-sample extension of graph adjacency spectral embedding. In *International Conference on Machine Learning*, pp. 2975–2984. PMLR, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the International Conference on Learning Representations*, 2018.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., and Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020a.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020b.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32:13153–13164, 2019.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Malinin, A. and Gales, M. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *Advances in Neural Information Processing Systems*, 2019.

McSherry, F. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pp. 529–537. IEEE, 2001.

Meinke, A. and Hein, M. Towards neural networks that provably know when they don't know. In *Proceedings of the International Conference on Learning Representations*, 2020.

Ming, Y. and Li, Y. How does fine-tuning impact out-of-distribution detection for vision-language models? *International Journal of Computer Vision*, 2023.

Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., and Li, Y. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, 2022a.

Ming, Y., Fan, Y., and Li, Y. POEM: out-of-distribution detection with posterior sampling. In *Proceedings of the International Conference on Machine Learning*, pp. 15650–15665, 2022b.

Ming, Y., Yin, H., and Li, Y. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022c.

Ming, Y., Sun, Y., Dia, O., and Li, Y. How to exploit hyperspherical embeddings for out-of-distribution detection? In *Proceedings of the International Conference on Learning Representations*, 2023.

Morteza, P. and Li, Y. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7831–7840, 2022.

Narasimhan, H., Menon, A. K., Jitkrittum, W., and Kumar, S. Learning to reject meets ood detection: Are all abstentions created equal? *arXiv preprint arXiv:2301.12386*, 2023.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Ng, A., Jordan, M., and Weiss, Y. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

Park, S., Mok, J., Jung, D., Lee, S., and Yoon, S. On the powerfulness of textual outlier exposure for visual ood detection. *arXiv preprint arXiv:2310.16492*, 2023.

Quispe, A. M., Petitjean, C., and Heutte, L. Extreme learning machine for out-of-sample extension in laplacian eigenmaps. *Pattern Recognition Letters*, 74:68–73, 2016.

Ren, J., Fort, S., Liu, J., Roy, A. G., Padhy, S., and Lakshminarayanan, B. A simple fix to mahalanobis distance for improving near-ood detection. *CoRR*, abs/2106.09022, 2021.

Ren, J., Luo, J., Zhao, Y., Krishna, K., Saleh, M., Lakshminarayanan, B., and Liu, P. J. Out-of-distribution detection and selective generation for conditional language models. In *International Conference on Learning Representations*, 2023.

Sehwag, V., Chiang, M., and Mittal, P. Ssd: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021a.

Sehwag, V., Chiang, M., and Mittal, P. Ssd: A unified framework for self-supervised outlier detection. In *Proceedings of the International Conference on Learning Representations*, 2021b.

Shaham, U., Stanton, K., Li, H., Nadler, B., Basri, R., and Kluger, Y. Spectralnet: Spectral clustering using deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Shen, K., Jones, R. M., Kumar, A., Xie, S. M., HaoChen, J. Z., Ma, T., and Liang, P. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 19847–19878, 2022.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the International Conference on Machine Learning*, pp. 20827–20840, 2022.

Sun, Y., Shi, Z., and Li, Y. A graph-theoretic framework for understanding open-world representation learning. In *Advances in Neural Information Processing Systems*, 2023a.

Sun, Y., Shi, Z., Liang, Y., and Li, Y. When and how does known class help discover unknown ones? provable

understandings through spectral analysis. In *International Conference on Machine Learning*, 2023b.

Sussman, D. L., Tang, M., Fishkind, D. E., and Priebe, C. E. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

Tack, J., Mo, S., Jeong, J., and Shin, J. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, 2020.

Tang, M., Sussman, D. L., and Priebe, C. E. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430, 2013.

Tao, L., Du, X., Zhu, X., and Li, Y. Non-parametric outlier synthesis. In *Proceedings of the International Conference on Learning Representations*, 2023.

Trosset, M. W. and Priebe, C. E. The out-of-sample problem for classical multidimensional scaling. *Computational statistics & data analysis*, 52(10):4635–4642, 2008.

Uppaal, R., Hu, J., and Li, Y. Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection. In *Annual Meeting of the Association for Computational Linguistics*, 2023.

van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of the International Conference on Machine Learning*, pp. 9690–9700, 2020.

Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Wang, H., Liu, W., Bocchieri, A., and Li, Y. Can multi-label classification networks know what they don't know? *Proceedings of the Advances in Neural Information Processing Systems*, 2021.

Wang, Q., Liu, F., Zhang, Y., Zhang, J., Gong, C., Liu, T., and Han, B. Watermarking for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:15545–15557, 2022a.

Wang, Q., Fang, Z., Zhang, Y., Liu, F., Li, Y., and Han, B. Learning to augment distributions for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2023a.

Wang, Q., Ye, J., Liu, F., Dai, Q., Kalander, M., Liu, T., HAO, J., and Han, B. Out-of-distribution detection with implicit outlier transformation. In *International Conference on Learning Representations*, 2023b.

Wang, Y., Zou, J., Lin, J., Ling, Q., Pan, Y., Yao, T., and Mei, T. Out-of-distribution detection via conditional kernel independence model. In *Advances in Neural Information Processing Systems*, 2022b.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., and Li, Y. Mitigating neural network overconfidence with logit normalization. In *Proceedings of the International Conference on Machine Learning*, pp. 23631–23644, 2022.

Wen, Y., Tran, D., and Ba, J. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020.

Wu, Q., Chen, Y., Yang, C., and Yan, J. Energy-based out-of-distribution detection for graph neural networks. In *International Conference on Learning Representations*, 2023.

Yang, J., Wang, H., Feng, L., Yan, X., Zheng, H., Zhang, W., and Liu, Z. Semantically coherent out-of-distribution detection. In *Proceedings of the International Conference on Computer Vision*, pp. 8281–8289, 2021a.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021b.

Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., PENG, W., Wang, H., Chen, G., Li, B., Sun, Y., Du, X., Zhou, K., Zhang, W., Hendrycks, D., Li, Y., and Liu, Z. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances on Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

Yang, P., Liang, J., Cao, J., and He, R. Auto: Adaptive outlier optimization for online test-time ood detection. *arXiv preprint arXiv:2303.12267*, 2023.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

Zheng, H., Wang, Q., Fang, Z., Xia, X., Liu, F., Liu, T., and Han, B. Out-of-distribution detection learning with unreliable out-of-distribution sources. *arXiv preprint arXiv:2311.03236*, 2023.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Zhou, Z., Guo, L.-Z., Cheng, Z., Li, Y.-F., and Pu, S. STEP: Out-of-distribution detection in the presence of limited in-distribution labeled data. In *Advances in Neural Information Processing Systems*, 2021.

Zhu, J., Li, H., Yao, J., Liu, T., Xu, J., and Han, B. Unleashing mask: Explore the intrinsic out-of-distribution detection capability. *arXiv preprint arXiv:2306.03715*, 2023a.

Zhu, J., Yu, G., Yao, J., Liu, T., Niu, G., Sugiyama, M., and Han, B. Diversified outlier exposure for out-of-distribution detection via informative extrapolation. *arXiv preprint arXiv:2310.13923*, 2023b.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pp. 912–919, 2003.

Zhu, Y., Chen, Y., Xie, C., Li, X., Zhang, R., Xue', H., Tian, X., bolun zheng, and Chen, Y. Boosting out-of-distribution detection with typical features. In *Advances in Neural Information Processing Systems*, 2022.

When and How Does In-Distribution Label Help Out-of-Distribution Detection? (Appendix)

A. Notations, Assumptions and Important Constants

Here we summarize the important notations and constants in Table 4, and restate necessary definitions and assumptions in Section A.2.

A.1. Notations

Please see Table 4 for detailed notations.

Table 4: Main notations and their descriptions.

Notation	Description
Spaces	
$\mathcal{X}, \mathcal{Z}, \mathcal{Y}$	the input, representation space and the label space.
Distributions	
$\mathbb{P}_{\mathrm{id}}, \mathbb{P}_{\mathrm{ood}}$	data distribution for ID data and OOD data
$\mathbb{P}_{\mathcal{XY}}$	the joint data distribution for ID data.
Data and Models	
$\mathbf{x}$	input to the neural network
$\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}$ and $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)}$	adjacency matrix between OOD and ID data in the labeled and unlabeled case
$\mathbf{A}^{(u)}$ and $\mathbf{A}^{(l)}$	adjacency matrix for ID data in the labeled and unlabeled case
$\tilde{\mathbf{A}}^{(u)}$ and $\tilde{\mathbf{A}}^{(l)}$	normalized adjacency matrix for ID data in the labeled and unlabeled case
$\mathbf{D}^{(u)}$ and $\mathbf{D}^{(l)}$	ID diagonal matrix with the diagonal elements being the row sum of $\mathbf{A}^{(u)}$ and $\mathbf{A}^{(l)}$.
$\mathbf{D}_{\mathrm{ood}}^{(u)}$ and $\mathbf{D}_{\mathrm{ood}}^{(l)}$	OOD diagonal matrix with the diagonal elements being the row sum of $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}$ and $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)}$.
$\mathcal{T}$	augmentation graph
$\mathbf{w}, \boldsymbol{\theta}$	weight of the ID feature extractor and the linear probing layer
$\mathbf{h_w}$ and $\mathbf{g_\theta}$	feature extractor on ID data and linear probing layer for OOD detection
$y, \mathbf{y}$	binary label for linear probing, vectorized one-hot label for y
$\mathbf{z}$	feature for single input $\mathbf{x}$
$\mathbf{Z}^{(u)}$ and $\mathbf{Z}^{(l)}$	representation matrix for ID data in the labeled and unlabeled case
$\mathbf{Z}_{\mathrm{ood}}^{(u)}$ and $\mathbf{Z}_{\mathrm{ood}}^{(l)}$	representation matrix for OOD data in the labeled and unlabeled case
$\mathbf{Z}_{\mathrm{all}}^{(u)}$ and $\mathbf{Z}_{\mathrm{all}}^{(l)}$	representation matrix for both ID and OOD data in the labeled and unlabeled case
$\mathfrak{q}$ and $\mathfrak{p}$	semantic connection from each ID/OOD data to the labeled ID data.
N, M	size of $\mathbf{Z}$, size of $\mathbf{Z}_{\mathrm{ood}}$
$\boldsymbol{\lambda}$	eigenvalue vector of $\tilde{\mathbf{A}}$
$\mathbf{v}_j$	the j-th eigenvector of $\tilde{\mathbf{A}}$
$\mathbf{V}_k, \mathbf{V}_\varnothing$	the first k eigenvectors of $\tilde{\mathbf{A}}$/the latter $N-k$ eigenvectors of $\tilde{\mathbf{A}}$ (null space of $\mathbf{V}_k$)
Distances	
$\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_F$	ℓ_1, ℓ_2 norm and Frobenius norm
Loss and Risk	
$\mathcal{L}_{\mathrm{unlabeled}}(\cdot), \mathcal{L}_{\mathrm{labeled}}(\cdot)$	ID loss function in the unlabeled and labeled case
$R(\mathbf{Z}_{\mathrm{all}})$	the empirical risk w.r.t. linear probing module over feature set $\mathbf{Z}_{\mathrm{all}}$
$\overline{R}(\mathbf{Z}_{\mathrm{all}})$	the upper bound of $R(\mathbf{Z}_{\mathrm{all}})$
$\mathcal{G}$	linear probing error difference with and without ID labels
Constants	
ϕ_u, ϕ_l	weight coefficients for the unlabeled and labeled case.
k	dimension of the feature representation
τ	the parameter that measures the eigengap of $\tilde{\mathbf{A}}^{(u)}$
r	the maximum ℓ_2 norm of the ID features, i.e., $r = \max_{\mathbf{z} \in \mathbf{Z}^{(u)}} \|\mathbf{z}\|_2$
C	constants in Theorem 1

A.2. Assumptions

Assumption 1 (Property of ID adjacency matrix)**.** We assume the adjacency matrix $\tilde{\mathbf{A}}^{(u)}$ has the following property: There exists a positive constant $k \leq N$ such that the k-th eigengap of $\tilde{\mathbf{A}}^{(u)}$ satisfies that $\boldsymbol{\lambda}_k^{(u)} \geq \tau \boldsymbol{\lambda}_{k+1}^{(u)}, \tau > k$, where $\boldsymbol{\lambda}_k^{(u)}$ is the k-th largest eigenvalue of matrix $\tilde{\mathbf{A}}^{(u)}$.

Remark 1. We have empirically verified our assumption using both simulated and real-world datasets in Section E.

Assumption 2 (Property of vectors that depict the semantic connection between each ID/OOD data to the labeled ID data)**.** Similar to (Sun et al., 2023a), we assume each row of the matrix q lies in the linear span of $\mathbf{V}_k^{(u)}$ and $\mathbf{V}_k^{(u)}[\boldsymbol{\Sigma}_k^{(u)}]^{-\frac{1}{2}}$, i.e., $\mathbf{V}_\varnothing^{(u)\top} \mathfrak{q}_i = 0, \mathbf{V}_k^{(u)} \mathbf{V}_k^{(u)\top} \mathfrak{q}_i = \mathfrak{q}_i$ and $\mathbf{V}_k^{(u)}[\boldsymbol{\Sigma}_k^{(u)}]^{-1} \mathbf{V}_k^{(u)\top} \mathfrak{q}_i = \mathfrak{q}_i$.

Remark 2. Following the assumptions made in the existing literature (Sun et al., 2023a), the assumption is used to simplify $\mathbf{V}_k^{(u)} \mathbf{V}_k^{(u)\top} \mathfrak{q}_i \mathfrak{q}_i^\top \mathbf{V}_\varnothing^{(u)} \mathbf{V}_\varnothing^{(u)\top} + \mathbf{V}_k^{(u)} \mathbf{V}_k^{(u)\top} \mathfrak{q}_i \mathfrak{q}_i^\top$ to $\mathfrak{q}_i \mathfrak{q}_i^\top$ and simplify $\mathbf{V}_k^{(u)}[\boldsymbol{\Sigma}_k^{(u)}]^{-1} \mathbf{V}_k^{(u)\top} \mathfrak{q}_i$ to $\mathfrak{q}_i$.

B. Main Theorems

In this section, we provide a detailed and formal version of our main theorems with a complete description of the constant terms and other additional details that are omitted in the main paper.

Theorem 1 (Lower bound of the error difference between unlabeled and labeled cases, recap of Theorem 1 in the main paper)**.** *Suppose we have adjacency matrices $\tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}^{(l)} \in \mathbb{R}^{N \times N}$ and $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}, \tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)} \in \mathbb{R}^{M \times N}$ for both the labeled and unlabeled cases. If Assumption 2 holds, given positive constants ϕ_l, C, the error difference $\mathcal{G}$ in Equation 11 is bounded by*

$$\mathcal{G} \geq \frac{C\phi_l}{N+M} \epsilon(\mathfrak{p}, \mathfrak{q}, \tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}), \tag{13}$$

where $\mathfrak{q} \in \mathbb{R}^{c \times N}$ with each row being defined as $(\mathfrak{q}_i)_{\mathbf{x}} = \mathbb{E}_{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}} \mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l), \mathbf{x} \in \mathcal{X}_{\mathrm{id}}$. Similarly, we have $\mathfrak{p} \in \mathbb{R}^{c \times M}$ and $(\mathfrak{p}_i)_{\mathbf{x}} = \mathbb{E}_{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}} \mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l), \mathbf{x} \in \mathcal{X}_{\mathrm{ood}}$. Furthermore,

$$\begin{aligned}
\epsilon(\mathfrak{p}, \mathfrak{q}, \tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}) = {} & 2 \sum_{i=1}^{c} \mathrm{Tr}\left(\mathfrak{p}_i \mathfrak{q}_i^\top \cdot \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) + \left(1 - \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2\right) \sum_{i=1}^{c} \|\mathfrak{q}_i\|_F^2 + \\
& r^2 \left(\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2 \frac{2(\tau - k)}{\tau - 1} - 2\right) \cdot \sum_{i=1}^{c} \|\mathfrak{q}_i\|_1,
\end{aligned} \tag{14}$$

where τ is a constant that measures the k-th spectral gap of matrix $\tilde{\mathbf{A}}^{(u)}$, i.e., $\boldsymbol{\lambda}_k^{(u)} \geq \tau \boldsymbol{\lambda}_{k+1}^{(u)}$ and $\boldsymbol{\lambda}_k^{(u)}$ is the k-th largest singular value of $\tilde{\mathbf{A}}^{(u)}$. r is the maximum l_2 norm of the ID representations, i.e., $r = \max_{\mathbf{z} \in \mathbf{Z}^{(u)}} \|\mathbf{z}\|_2$.

Theorem 2 (Simplified version of Theorem 1, recap of Theorem 2 in the main paper)**.** *Under the same conditions in Theorem 1, if the number of ID and OOD samples N and M is sufficiently large, assume the k-th spectral gap of $\tilde{\mathbf{A}}^{(u)}$ satisfies $\tau > k$ (Assumption 1 holds), then the main error component $\epsilon(\mathfrak{p}, \mathfrak{q}, \tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)})$ in Equation 14 satisfies*

$$\epsilon(\mathfrak{p}, \mathfrak{q}, \tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}) \geq \left[1 + \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \left(2N^2 - \|\tilde{\mathbf{A}}^{(u)}\|_F^2\right)\right] \sum_{i=1}^{c} \|\mathfrak{q}_i\|_F^2. \tag{15}$$

C. Proofs of Main Theorems

C.1. Proof of Theorem 1

Before proving Theorem 1, we first explain the framework introduced in Section 4 of the main paper to analyze the difference in the adjacency matrices and the corresponding feature representations between the unlabeled and labeled cases. Specifically, we proposed to analyze the adjacency matrix in the labeled case by perturbation analysis.

Matrix perturbation in the labeled case. Recall that we define in Definition 4 that the adjacency matrix in the labeled case (l) is the unlabeled one (u) plus connectivity incurred by the ID labels, which can be regarded as the perturbation of the labeling information. Therefore, for the ID adjacency matrix with ID labels, we have the perturbation on $\mathbf{A}^{(u)}$ as follows:

$$\mathbf{A}^{(l)} = \phi_u \mathbf{A}^{(u)} + \phi_l \Delta \mathbf{A}, \tag{16}$$

where $\Delta \mathbf{A} \in \mathbb{R}^{N \times N}$ can be calculated based on the augmentation graph $\mathcal{T}$ according to Definition 4. Following (Sun et al., 2023a), we study the perturbation from two aspects: (1) The direction of the perturbation which is given by $\Delta \mathbf{A}$, (2) The perturbation magnitude ϕ_l. We first consider the perturbation direction $\Delta \mathbf{A}$ and recall that we defined the concrete form in Definition 4:

$$[\Delta \mathbf{A}]_{\mathbf{x}\mathbf{x}'} \triangleq \sum_{i \in \mathcal{Y}} \mathbb{E}_{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}} \mathbb{E}_{\bar{\mathbf{x}}'_l \sim \mathbb{P}_{l_i}} \mathcal{T}(\mathbf{x} \mid \bar{\mathbf{x}}_l) \mathcal{T}(\mathbf{x}' \mid \bar{\mathbf{x}}'_l), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}_{\text{id}} \tag{17}$$

where $\mathcal{T}$ is the augmentation graph. In our theory, we consider $\|\mathcal{Y}\| = c$, and then we observe that $\Delta \mathbf{A}$ is the sum of c rank-1 matrices, which can be written as:

$$\Delta \mathbf{A} = \sum_{i=1}^{c} \mathfrak{q}_i \mathfrak{q}_i^\top, \tag{18}$$

where $\mathfrak{q} \in \mathbb{R}^{N \times c}$ with each column defined as $(\mathfrak{q}_i)_{\mathbf{x}} = \mathbb{E}_{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}} \mathcal{T}(\mathbf{x} \mid \bar{\mathbf{x}}_l), \mathbf{x} \in \mathcal{X}_{\text{id}}$ and $\bar{\mathbf{x}}_l$ is the labeled ID data with label i. And following (Sun et al., 2023a), we define the diagonal matrix $\mathbf{D}^{(l)}$ as follows:

$$\mathbf{D}^{(l)} = \phi_u \cdot \mathbf{D}^{(u)} + \sum_{i=1}^{c} \phi_l \cdot \text{diag}(\mathfrak{q}_i). \tag{19}$$

Without losing the generality, we let $\phi_u \cdot \text{diag}(\mathbf{1}_N^\top \mathbf{A}^{(u)}) = \mathbf{I}_{N \times N}$, which means the ID nodes have equal degree in the unlabeled case. We then have:

$$\mathbf{D}^{(l)} \triangleq \text{diag}\left(\mathbf{1}_N^\top \mathbf{A}^{(l)}\right) = \mathbf{I}_{N \times N} + \phi_l \cdot \sum_{i=1}^{c} \text{diag}(\mathfrak{q}_i). \tag{20}$$

The perturbation function of representation. We then consider a more generalized form for the ID adjacency matrix $\mathbf{A}$:

$$\mathbf{A}(\phi_l) = \phi_u \mathbf{A}^{(u)} + \phi_l \cdot \sum_{i=1}^{c} \mathfrak{q}_i \mathfrak{q}_i^\top. \tag{21}$$

For the perturbation on the OOD-ID adjacency matrix $\tilde{\mathbf{A}}_{\text{OI}}$, we define $\mathfrak{p} \in \mathbb{R}^{M \times c}$ with each column defined as $(\mathfrak{p}_i)_{\mathbf{x}} = \mathbb{E}_{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}} \mathcal{T}(\mathbf{x} \mid \bar{\mathbf{x}}_l), \mathbf{x} \sim \mathcal{X}_{\text{ood}}$ and $\bar{\mathbf{x}}_l$ is the labeled ID data with label l_i. Similar to Equation 17, we can define the concrete perturbation formula for $\tilde{\mathbf{A}}_{\text{OI}}$ as follows:

$$[\Delta \tilde{\mathbf{A}}_{\text{OI}}]_{\mathbf{x}\mathbf{x}'} \triangleq \sum_{i \in \mathcal{Y}} \mathbb{E}_{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}} \mathbb{E}_{\bar{\mathbf{x}}'_l \sim \mathbb{P}_{l_i}} \mathcal{T}(\mathbf{x} \mid \bar{\mathbf{x}}_l) \mathcal{T}(\mathbf{x}' \mid \bar{\mathbf{x}}'_l), \quad \mathbf{x}' \sim \mathcal{X}_{\text{id}}, \mathbf{x} \sim \mathcal{X}_{\text{ood}}. \tag{22}$$

And we have:

$$\tilde{\mathbf{A}}_{\text{OI}}(\phi_l) = \phi_u \tilde{\mathbf{A}}_{\text{OI}}^{(u)} + \phi_l \cdot \sum_{i=1}^{c} \mathfrak{p}_i \mathfrak{q}_i^\top. \tag{23}$$

17

Similarly, for the ID diagonal matrix $\mathbf{D}$, we have that:

$$\mathbf{D}(\phi_l) = \mathbf{I}_{N \times N} + \phi_l \cdot \sum_{i=1}^{c} \operatorname{diag}(\mathfrak{q}_i). \tag{24}$$

In Equations 21 and 23, we treat the adjacency matrix as a function of the "labeling perturbation" magnitude ϕ_l. It is clear that $\mathbf{A}(0) = \phi_u \mathbf{A}^{(u)}$ and $\tilde{\mathbf{A}}_{\mathrm{OI}}(0) = \phi_u \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}$ which are the scaled adjacency matrices in the unlabeled case. When we let the adjacency matrix be a function of ϕ_l, the normalized form, and the derived feature representation should also be the function of ϕ_l. We proceed by defining these terms.

The normalized ID adjacency matrix is given by:

$$\tilde{\mathbf{A}}(\phi_l) = \mathbf{D}(\phi_l)^{-\frac{1}{2}} \mathbf{A}(\phi_l) \mathbf{D}(\phi_l)^{-\frac{1}{2}}, \tag{25}$$

In addition, for in-distribution feature representation $\mathbf{Z}(\phi_l)$, it is derived from the top-k eigen components of $\tilde{\mathbf{A}}(\phi_l)$. Specifically, we have:

$$\mathbf{Z}(\phi_l)\mathbf{Z}(\phi_l)^{\top} = \mathbf{D}(\phi_l)^{-\frac{1}{2}} \tilde{\mathbf{A}}_k(\phi_l) \mathbf{D}(\phi_l)^{-\frac{1}{2}} = \mathbf{D}(\phi_l)^{-\frac{1}{2}} \sum_{j=1}^{k} \boldsymbol{\lambda}_j(\phi_l) \boldsymbol{\Phi}_j(\phi_l) \mathbf{D}(\phi_l)^{-\frac{1}{2}}, \tag{26}$$

where $\tilde{\mathbf{A}}_k(\phi_l)$ is the top-k SVD components of $\tilde{\mathbf{A}}(\phi_l)$ and can be written as $\tilde{\mathbf{A}}_k(\phi_l) = \sum_{j=1}^{k} \boldsymbol{\lambda}_j(\phi_l) \boldsymbol{\Phi}_j(\phi_l)$. Here $\boldsymbol{\lambda}_j(\phi_l)$ and $\boldsymbol{\Phi}_j(\phi_l) = \mathbf{v}_j(\phi_l)\mathbf{v}_j^{\top}(\phi_l)$ are the j-th eigenvalue and eigen projector of matrix $\tilde{\mathbf{A}}(\phi_l)$. For simplicity, when $\phi_l = 0$, we remove the suffix (0) and give the following definitions:

$$\tilde{\mathbf{A}}(0) = \tilde{\mathbf{A}}^{(u)}, \mathbf{Z}(0) = \mathbf{Z}^{(u)}, \mathbf{Z}_{\mathrm{ood}}(0) = \mathbf{Z}_{\mathrm{ood}}^{(u)}, \mathbf{Z}_{\mathrm{all}}(0) = \mathbf{Z}_{\mathrm{all}}^{(u)}, \boldsymbol{\lambda}_j(0) = \boldsymbol{\lambda}_j^{(u)}, \mathbf{v}_j(0) = \mathbf{v}_j^{(u)}, \boldsymbol{\Phi}_j(0) = \boldsymbol{\Phi}_j^{(u)}. \tag{27}$$

We proceed to provide the five concrete steps to prove Theorem 1.

Step 1. Recall that our analysis is based on the upper bound of the linear probing loss, and focuses on the error difference between the labeled and the unlabeled case, which is formulated in Equation 11 of the main paper. In the context of perturbation analysis, we can reformulate the error difference $\mathcal{G}$ as a function of ϕ_l as follows:

$$\mathcal{G}(\phi_l) = \overline{R}(\mathbf{Z}_{\mathrm{all}}(0)) - \overline{R}(\mathbf{Z}_{\mathrm{all}}(\phi_l)). \tag{28}$$

By the definition of the derivative, we can rewrite Equation 28:

$$\mathcal{G}(\phi_l) = - \left. \frac{\mathrm{d}\overline{R}\left(\mathbf{Z}_{\mathrm{all}}(\phi_l)\right)}{\mathrm{d}\phi_l} \right|_{\phi_l=0} \cdot \phi_l \tag{29}$$

With Lemma 3, we can get the lower bound of $\mathcal{G}(\phi_l)$ as follows:

$$\mathcal{G}(\phi_l) \geq \frac{2}{3(N+M)} \phi_l \cdot \max \lambda \left(\mathbf{Z}_{\mathrm{all}}(\phi_l)^{\top}\mathbf{Z}_{\mathrm{all}}(\phi_l)\right) \cdot \operatorname{Tr}\left(\left(\mathbf{Z}_{\mathrm{all}}(\phi_l)\mathbf{Z}_{\mathrm{all}}(\phi_l)^{\top}\right)' \mathbf{y}\mathbf{y}^{\top}\right) \Big|_{\phi_l=0} \tag{30}$$

According to Lemma 4, $\max \lambda \left(\mathbf{Z}_{\mathrm{all}}(\phi_l)^{\top}\mathbf{Z}_{\mathrm{all}}(\phi_l)\right)\big|_{\phi_l=0} > 0$ is a positive constant. Here $\max \lambda \left(\mathbf{Z}_{\mathrm{all}}(\phi_l)^{\top}\mathbf{Z}_{\mathrm{all}}(\phi_l)\right)$ means the maximum eigenvalue of the matrix $\mathbf{Z}_{\mathrm{all}}(\phi_l)^{\top}\mathbf{Z}_{\mathrm{all}}(\phi_l)$. We then proceed to analyze the key component $\operatorname{Tr}\left(\left(\mathbf{Z}_{\mathrm{all}}(\phi_l)\mathbf{Z}_{\mathrm{all}}(\phi_l)^{\top}\right)' \mathbf{y}\mathbf{y}^{\top}\right)\big|_{\phi_l=0}$.

Next, according to Lemma 3.1 in (HaoChen et al., 2021), which implies that multiplying any invertible matrix by the features of the linear probing module does not change the linear probing performance. Therefore, we simplify the mathematical representation of the ID and OOD features in the Equations 6 and 8 in the main paper as follows by removing the terms $\mathbf{D}(\phi_l)$ and $\mathbf{D}_{\mathrm{ood}}(\phi_l)$:

$$\mathbf{Z}^{(l)} = \mathbf{V}_k^{(l)}[\boldsymbol{\Sigma}_k^{(l)}]^{\frac{1}{2}}, \quad \mathbf{Z}_{\mathrm{ood}}^{(l)} = \tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)}\mathbf{V}_k^{(l)}[\boldsymbol{\Sigma}_k^{(l)}]^{-\frac{1}{2}}, \tag{31}$$

which simplifies the notation and will be used in our later analysis.

Step 2. Recall that the input feature for linear probing $\mathbf{Z}(\phi_l)$ is defined as:

$$\mathbf{Z}_{\text{all}}(\phi_l) = \begin{pmatrix} \mathbf{Z}(\phi_l) \\ \mathbf{Z}_{\text{ood}}(\phi_l) \end{pmatrix} = \begin{pmatrix} \mathbf{V}_k(\phi_l)\sqrt{\mathbf{\Sigma}_k(\phi_l)} \\ \tilde{\mathbf{A}}_{\text{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\sqrt{\mathbf{\Sigma}_k(\phi_l)}^{-1} \end{pmatrix} \tag{32}$$

Therefore, it is natural to have that:

$$\begin{aligned}
\text{Tr}((\mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\top)'\mathbf{y}\mathbf{y}^\top)\Big|_{\phi_l=0} &= \text{Tr}\left(\left(\left(\begin{matrix} \mathbf{V}_k(\phi_l)\sqrt{\mathbf{\Sigma}_k(\phi_l)} \\ \tilde{\mathbf{A}}_{\text{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\sqrt{\mathbf{\Sigma}_k(\phi_l)}^{-1} \end{matrix}\right) \cdot \left(\sqrt{\mathbf{\Sigma}_k(\phi_l)}\mathbf{V}_k(\phi_l)^\top \quad \sqrt{\mathbf{\Sigma}_k(\phi_l)}^{-1}\mathbf{V}_k(\phi_l)^\top\tilde{\mathbf{A}}_{\text{OI}}^k(\phi_l)^\top\right)\right)'\mathbf{y}\mathbf{y}^\top\right)\Big|_{\phi_l=0} \\
&\geq \text{Tr}\left(\left(\begin{matrix} \mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^\top & \mathbf{V}_k(\phi_l)\mathbf{V}_k(\phi_l)^\top\tilde{\mathbf{A}}_{\text{OI}}(\phi_l) \\ \tilde{\mathbf{A}}_{\text{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\mathbf{V}_k(\phi_l)^\top & \tilde{\mathbf{A}}_{\text{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)^{-1}\mathbf{V}_k(\phi_l)^\top\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)^\top \end{matrix}\right)'\right)\Big|_{\phi_l=0} \cdot \text{Tr}(\mathbf{y}\mathbf{y}^\top) \\
&= \left(\text{Tr}\left(\left(\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^\top\right)'\right) + \text{Tr}\left(\left(\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)^{-1}\mathbf{V}_k(\phi_l)^\top\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)^\top\right)'\right)\Big|_{\phi_l=0}\right) \cdot \|\mathbf{y}\|_F^2
\end{aligned} \tag{33}$$

In the following steps, we focus on the lower bounds of the two separate terms, i.e., $\text{Tr}\left(\left(\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)^{-1}\mathbf{V}_k(\phi_l)^\top\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)^\top\right)'\right)\Big|_{\phi_l=0}$ and $\text{Tr}\left(\left(\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^\top\right)'\right)\Big|_{\phi_l=0}$, respectively.

Step 3. As in Lemma 5, we provide the lower bound for the first term $\text{Tr}\left(\left(\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^\top\right)'\right)\Big|_{\phi_l=0}$ as follows:

$$\text{Tr}\left(\left(\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^\top\right)'\right)\Big|_{\phi_l=0} \geq \sum_{i=1}^c \|\mathfrak{q}_i\|_F^2 - 2r^2\sum_{i=1}^c \|\mathfrak{q}_i\|_1, \tag{34}$$

where r denotes the maximum ℓ_2 norm of the ID features, i.e., $r = \max_{\mathbf{z}\in\mathbf{Z}^{(u)}} \|\mathbf{z}\|_2$, and $\mathfrak{q} \in \mathbb{R}^{N\times c}$ with each column defined as $(\mathfrak{q}_i)_\mathbf{x} = \mathbb{E}_{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l), \mathbf{x}\in\mathcal{X}_{\text{id}}$.

Step 4. We provide the lower bound for second term $\text{Tr}\left(\left(\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)^{-1}\mathbf{V}_k(\phi_l)^\top\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)^\top\right)'\right)\Big|_{\phi_l=0}$ as in Lemma 6 as follows:

$$\begin{aligned}
&\text{Tr}\left(\left(\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)^{-1}\mathbf{V}_k(\phi_l)^\top\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)^\top\right)'\right)\Big|_{\phi_l=0} \\
&\geq 2\sum_{i=1}^c \text{Tr}\left(\mathfrak{p}_i\mathfrak{q}_i^\top \cdot \tilde{\mathbf{A}}_{\text{OI}}^{(u)\top}\right) - \|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F^2\|\tilde{\mathbf{A}}^{(u)}\|_F^2 \cdot \left(\sum_{i=1}^c \|\mathfrak{q}_i\|_F^2 - \frac{2r^2(\tau-k)}{\tau-1}\right)\sum_{i=1}^c \|\mathfrak{q}_i\|_1),
\end{aligned} \tag{35}$$

where $\mathfrak{p} \in \mathbb{R}^{M\times c}$ is defined as $(\mathfrak{p}_i)_\mathbf{x} = \mathbb{E}_{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l), \mathbf{x}\in\mathcal{X}_{\text{ood}}$.

Step 5. Putting the result in Equations 34 and 35 together, if we denote $C = \frac{2}{3}\|\mathbf{y}\|_F^2 \cdot \max\lambda\left(\mathbf{Z}_{\text{all}}(0)^\top\mathbf{Z}_{\text{all}}(0)\right)$, we can get:

$$\begin{aligned}
\mathcal{G}(\phi_l) &\geq \frac{C\phi_l}{N+M} \cdot \text{Tr}\left(\left(\mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\top\right)'\right)\Big|_{\phi_l=0} \\
&\geq \frac{C\phi_l}{N+M}\left[2\sum_{i=1}^c \text{Tr}\left(\mathfrak{p}_i\mathfrak{q}_i^\top \cdot \tilde{\mathbf{A}}_{\text{OI}}^{(u)\top}\right) + \left(1 - \|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F^2 \cdot \|\tilde{\mathbf{A}}^{(u)}\|_F^2\right)\sum_{i=1}^c \|\mathfrak{q}_i\|_F^2 - r^2\left(2 - \|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F^2 \cdot \|\tilde{\mathbf{A}}^{(u)}\|_F^2\frac{2(\tau-k)}{\tau-1}\right) \cdot \sum_{i=1}^c \|\mathfrak{q}_i\|_1\right].
\end{aligned} \tag{36}$$

We have completed the proof of Theorem 1.

C.2. Proof of Theorem 2

As in Assumption 1, we can find a k such that $\tau > k$. Based on this condition, we can get the following

$$
\begin{aligned}
\tau > k &= 1 + k - 1 \\
&= 1 + \frac{\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2}{\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2 - 1}(k-1) - O\left(\frac{k-1}{\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2 - 1}\right).
\end{aligned}
\tag{37}
$$

When we have a sufficient number of ID or OOD data, $\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2$ will be large such that we can omit the term $O\left(\frac{k-1}{\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2 - 1}\right)$ because $k \le N$. Therefore, we will have the inequality of

$$
\tau > 1 + \frac{\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2}{\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2 - 1}(k-1).
\tag{38}
$$

Simply the above inequality, we can get that

$$
\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2 > \frac{\tau - 1}{\tau - k}.
\tag{39}
$$

Therefore we have

$$
-r^2 \left(2 - \left\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\right\|_F^2 \left\|\tilde{\mathbf{A}}^{(u)}\right\|_F^2 \frac{2(\tau-k)}{\tau-1}\right) \cdot \sum_{i=1}^{c} \|\mathfrak{q}_i\|_1 > 0,
\tag{40}
$$

Under this condition, the main error term $\epsilon(\mathfrak{p}, \mathfrak{q}, \tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)})$ in Theorem 1 can be further bounded by:

$$
\epsilon(\mathfrak{p}, \mathfrak{q}, \tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}) \ge 2 \sum_{i=1}^{c} \mathrm{Tr}\left(\mathfrak{p}_i \mathfrak{q}_i^\top \cdot \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) + \left(1 - \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2\right) \sum_{i=1}^{c} \|\mathfrak{q}_i\|_F^2,
\tag{41}
$$

Following the same analysis approach in literature (See Section 4.1 of (HaoChen et al., 2021)), we decompose the vectors $\mathfrak{p}$ and $\mathfrak{q}$ into numbers in each dimension instead of using the vectorized form for calculation. Specifically, we let $\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}) = \kappa, \mathbf{x}, \bar{\mathbf{x}} \sim \mathcal{X}_{\mathrm{ood}}$ and $\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}) = \gamma, \mathbf{x}, \bar{\mathbf{x}} \sim \mathcal{X}_{\mathrm{id}}$ Based on the adjacency matrix and its relationship with the augmentation graph (Equation 1), we can prove that the following result holds:

$$
\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)} = \sum_{i=1}^{c} \phi_u \frac{\gamma N}{M + N} \cdot [\mathfrak{p}_i, ..., \mathfrak{p}_i] + O(\kappa) = \sum_{i=1}^{c} \frac{1}{N^3 \gamma} \cdot [\mathfrak{p}_i, ..., \mathfrak{p}_i] + O(\kappa),
\tag{42}
$$

where $\phi_u = \frac{M+N}{N^4 \gamma^2}$. Then, we obtain that:

$$
\mathrm{Tr}\left(\mathfrak{p}_i \mathfrak{q}_i^\top \cdot \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) = \mathrm{Tr}\left(\mathfrak{q}_i^\top \cdot \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top} \cdot \mathfrak{p}_i\right) = \sum_{i=1}^{c} \gamma N^2 \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\mathfrak{q}_i\|_1 + O(\kappa).
\tag{43}
$$

After that, we can get:

$$
\epsilon(\mathfrak{p}, \mathfrak{q}, \tilde{\mathbf{A}}^{(u)}, \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}) \ge \left(1 - \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2\right) \sum_{i=1}^{c} \|\mathfrak{q}_i\|_F^2 + 2\gamma N^2 \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \sum_{i=1}^{c} \|\mathfrak{q}_i\|_1 + O(\kappa),
\tag{44}
$$

Further simplify the right-hand side of the above inequality, we can get:

$$
\begin{aligned}
&\left(1 - \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2\right) \sum_{i=1}^{c} \|\mathbf{q}_i\|_F^2 + 2\gamma N^2 \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \sum_{i=1}^{c} \|\mathbf{q}_i\|_1 + O(\kappa) \\
&= \left(1 - \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2\right) \sum_{i=1}^{c} \|\mathbf{q}_i\|_F^2 + 2c\gamma N^2 \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 N\gamma + O(\kappa) \\
&= \left(1 - \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2\right) \sum_{i=1}^{c} \|\mathbf{q}_i\|_F^2 + 2N^2 \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \cdot \sum_{i=1}^{c} \|\mathbf{q}_i\|_F^2 + O(\kappa) \\
&= \left(1 - \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \|\tilde{\mathbf{A}}^{(u)}\|_F^2 + 2N^2 \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2\right) \sum_{i=1}^{c} \|\mathbf{q}_i\|_F^2 + O(\kappa)
\end{aligned}
\tag{45}
$$

We have completed the proof of Theorem 2.

D. Necessary Lemmas

D.1. Details for Spectral Contrastive Learning

Lemma 1. *(Recap of Lemma 1 in the main paper) We define* $\mathbf{f_x} = \sqrt{\zeta_\mathbf{x}}\mathbf{h_w}(\mathbf{x})$ *for some function* $\mathbf{h_w}$. *Recall* ϕ_u, ϕ_l *are two weight coefficients given in Definition 4, then minimizing the loss function* $\mathcal{L}(\mathbf{F}^{(l)}, \mathbf{A}^{(l)})$ *in Equation 3 is equivalent to minimizing the surrogate loss in Equation 46.*

$$
\begin{aligned}
\mathcal{L}_{\text{labeled}}(\mathbf{h_w}) \triangleq &-2\phi_l \mathcal{L}_1(\mathbf{h_w}) - 2\phi_u \mathcal{L}_2(\mathbf{h_w}) + \\
&\phi_l^2 \mathcal{L}_3(\mathbf{h_w}) + 2\phi_l\phi_u \mathcal{L}_4(\mathbf{h_w}) + \phi_u^2 \mathcal{L}_5(\mathbf{h_w}),
\end{aligned}
\tag{46}
$$

where

$$
\begin{aligned}
\mathcal{L}_1(\mathbf{h_w}) &= \sum_{i\in\mathcal{Y}} \mathop{\mathbb{E}}_{\substack{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i},\bar{\mathbf{x}}_l'\sim\mathbb{P}_{l_i},\\ \mathbf{x}\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_l),\mathbf{x}^+\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_l')}} \left[\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}^+\right)\right], \\
\mathcal{L}_2(\mathbf{h_w}) &= \mathop{\mathbb{E}}_{\substack{\bar{\mathbf{x}}_u\sim\mathbb{P}_{\text{id}},\\ \mathbf{x}\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_u),\mathbf{x}^+\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_u)}} \left[\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}^+\right)\right], \\
\mathcal{L}_3(\mathbf{h_w}) &= \sum_{i,j\in\mathcal{Y}} \mathop{\mathbb{E}}_{\substack{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i},\bar{\mathbf{x}}_l'\sim\mathbb{P}_{l_j},\\ \mathbf{x}\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_l),\mathbf{x}^-\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_l')}} \left[\left(\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\right], \\
\mathcal{L}_4(\mathbf{h_w}) &= \sum_{i\in\mathcal{Y}} \mathop{\mathbb{E}}_{\substack{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i},\bar{\mathbf{x}}_u\sim\mathbb{P}_{\text{id}},\\ \mathbf{x}\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_l),\mathbf{x}^-\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_u)}} \left[\left(\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\right], \\
\mathcal{L}_5(\mathbf{h_w}) &= \mathop{\mathbb{E}}_{\substack{\bar{\mathbf{x}}_u\sim\mathbb{P}_{\text{id}},\bar{\mathbf{x}}_u'\sim\mathbb{P}_{\text{id}},\\ \mathbf{x}\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_u),\mathbf{x}^-\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_u')}} \left[\left(\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\right].
\end{aligned}
\tag{47}
$$

Proof. We can expand $\mathcal{L}_{\text{labeled}}(\mathbf{h_w})$ and obtain

$$
\begin{aligned}
\mathcal{L}_{\text{labeled}}(\mathbf{h_w}) &= \sum_{\mathbf{x},\mathbf{x}'\in\mathcal{X}_{\text{id}}} \left(\frac{\zeta_{\mathbf{xx}'}}{\sqrt{\zeta_\mathbf{x}\zeta_{\mathbf{x}'}}} - \mathbf{f_x}^\top \mathbf{f}_{\mathbf{x}'}\right)^2 = \\
\text{const} &+ \sum_{\mathbf{x},\mathbf{x}'\in\mathcal{X}_{\text{id}}} \left(-2\zeta_{\mathbf{xx}'}\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}'\right) + \zeta_\mathbf{x}\zeta_{\mathbf{x}'}\left(\mathbf{h_w}(\mathbf{x})^\top \mathbf{h_w}\left(\mathbf{x}'\right)\right)^2\right),
\end{aligned}
\tag{48}
$$

where $\mathbf{f_x} = \sqrt{\zeta_\mathbf{x}}\mathbf{h_w}(\mathbf{x})$ is a re-scaled version of $\mathbf{h_w}(\mathbf{x})$. At a high level, we follow the proof in (HaoChen et al., 2021), while the specific form of loss varies with the different definitions of positive/negative pairs. The form of $\mathcal{L}_{\text{labeled}}(\mathbf{h_w})$ is derived from plugging $\zeta_{\mathbf{xx}'}$ and $\zeta_\mathbf{x}$.

Recall that $\zeta_{\mathbf{xx}'}$ is defined by

$$
\zeta_{\mathbf{xx}'} = \phi_l \sum_{i\in\mathcal{Y}} \mathbb{E}_{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i}}\mathbb{E}_{\bar{\mathbf{x}}_l'\sim\mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l)\mathcal{T}\left(\mathbf{x}'|\bar{\mathbf{x}}_l'\right) + \phi_u\mathbb{E}_{\bar{\mathbf{x}}_u\sim\mathbb{P}_{\text{id}}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_u)\mathcal{T}\left(\mathbf{x}'|\bar{\mathbf{x}}_u\right),
$$

and $\zeta_\mathbf{x}$ is given by

$$
\begin{aligned}
\zeta_\mathbf{x} &= \sum_{\mathbf{x}'} \zeta_{\mathbf{xx}'} \\
&= \phi_l \sum_{i\in\mathcal{Y}} \mathbb{E}_{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i}}\mathbb{E}_{\bar{\mathbf{x}}_l'\sim\mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l)\sum_{\mathbf{x}'}\mathcal{T}\left(\mathbf{x}'|\bar{\mathbf{x}}_l'\right) + \phi_u\mathbb{E}_{\bar{\mathbf{x}}_u\sim\mathbb{P}_{\text{id}}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_u)\sum_{x'}\mathcal{T}\left(\mathbf{x}'|\bar{\mathbf{x}}_u\right) \\
&= \phi_l \sum_{i\in\mathcal{Y}} \mathbb{E}_{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l) + \phi_u\mathbb{E}_{\bar{\mathbf{x}}_u\sim\mathbb{P}_{\text{id}}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_u).
\end{aligned}
$$

Plugging in $\zeta_{\mathbf{x}\mathbf{x}'}$ we have,

$$
\begin{aligned}
&-2\sum_{\mathbf{x},\mathbf{x}'\in\mathcal{X}_{\mathrm{id}}}\zeta_{\mathbf{x}\mathbf{x}'}\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}'\right)\\
=&-2\sum_{\mathbf{x},\mathbf{x}^+\in\mathcal{X}_{\mathrm{id}}}\zeta_{\mathbf{x}\mathbf{x}^+}\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}^+\right)\\
=&-2\phi_l\sum_{i\in\mathcal{Y}}\mathbb{E}_{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i}}\mathbb{E}_{\bar{\mathbf{x}}_l'\sim\mathbb{P}_{l_i}}\sum_{\mathbf{x},\mathbf{x}'\in\mathcal{X}_{\mathrm{id}}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l)\mathcal{T}\left(\mathbf{x}'|\bar{\mathbf{x}}_l'\right)\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}'\right)\\
&-2\phi_u\mathbb{E}_{\bar{\mathbf{x}}_u\sim\mathbb{P}_{\mathrm{id}}}\sum_{\mathbf{x},\mathbf{x}'}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_u)\mathcal{T}\left(\mathbf{x}'|\bar{\mathbf{x}}_u\right)\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}'\right)\\
=&-2\phi_l\sum_{i\in\mathcal{Y}}\mathbb{E}_{\substack{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i},\bar{\mathbf{x}}_l'\sim\mathbb{P}_{l_i},\\\mathbf{x}\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_l),\mathbf{x}^+\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_l')}}\left[\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}^+\right)\right]\\
&-2\phi_u\mathbb{E}_{\substack{\bar{\mathbf{x}}_u\sim\mathbb{P}_{\mathrm{id}},\\\mathbf{x}\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_u),\mathbf{x}^+\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_u)}}\left[\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}^+\right)\right]\\
=&-2\phi_l\mathcal{L}_1(\mathbf{h_w})-2\phi_u\mathcal{L}_2(\mathbf{h_w}).
\end{aligned}
$$

Plugging $\zeta_{\mathbf{x}}$ and $\zeta_{\mathbf{x}'}$ we have,

$$
\begin{aligned}
&\sum_{\mathbf{x},\mathbf{x}'\in\mathcal{X}_{\mathrm{id}}}\zeta_{\mathbf{x}}\zeta_{\mathbf{x}'}\left(\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}'\right)\right)^2\\
=&\sum_{\mathbf{x},\mathbf{x}^-\in\mathcal{X}_{\mathrm{id}}}\zeta_{\mathbf{x}}\zeta_{\mathbf{x}^-}\left(\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\\
=&\sum_{\mathbf{x},\mathbf{x}'\in\mathcal{X}_{\mathrm{id}}}\left(\phi_l\sum_{i\in\mathcal{Y}}\mathbb{E}_{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l)+\phi_u\mathbb{E}_{\bar{\mathbf{x}}_u\sim\mathbb{P}_{\mathrm{id}}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_u)\right)\\
&\quad\cdot\left(\phi_l\sum_{j\in\mathcal{Y}}\mathbb{E}_{\bar{\mathbf{x}}_l'\sim\mathbb{P}_{l_j}}\mathcal{T}(\mathbf{x}^-|\bar{\mathbf{x}}_l')+\phi_u\mathbb{E}_{\bar{\mathbf{x}}_u'\sim\mathbb{P}_{\mathrm{id}}}\mathcal{T}(\mathbf{x}^-|\bar{\mathbf{x}}_u')\right)\left(\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\\
=&\phi_l^2\sum_{\mathbf{x},\mathbf{x}^-\in\mathcal{X}_{\mathrm{id}}}\sum_{i\in\mathcal{Y}}\mathbb{E}_{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l)\sum_{j\in\mathcal{Y}}\mathbb{E}_{\bar{\mathbf{x}}_l'\sim\mathbb{P}_{l_j}}\mathcal{T}(\mathbf{x}^-|\bar{\mathbf{x}}_l')\left(\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\\
&+2\phi_u\phi_l\sum_{\mathbf{x},\mathbf{x}^-\in\mathcal{X}_{\mathrm{id}}}\sum_{i\in\mathcal{Y}}\mathbb{E}_{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l)\mathbb{E}_{\bar{\mathbf{x}}_u\sim\mathbb{P}_{\mathrm{id}}}\mathcal{T}(\mathbf{x}^-|\bar{\mathbf{x}}_u)\left(\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\\
&+\phi_u^2\sum_{\mathbf{x},\mathbf{x}^-\in\mathcal{X}_{\mathrm{id}}}\mathbb{E}_{\bar{\mathbf{x}}_u\sim\mathbb{P}_{\mathrm{id}}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_u)\mathbb{E}_{\bar{\mathbf{x}}_u'\sim\mathbb{P}_{\mathrm{id}}}\mathcal{T}(\mathbf{x}^-|\bar{\mathbf{x}}_u')\left(\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\\
=&\phi_l^2\sum_{i\in\mathcal{Y}}\sum_{j\in\mathcal{Y}}\mathbb{E}_{\substack{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i},\bar{\mathbf{x}}_l'\sim\mathbb{P}_{l_j},\\\mathbf{x}\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_l),\mathbf{x}^-\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_l')}}\left[\left(\mathbf{h_w}(\mathbf{x})^\top\mathbf{h_w}\left(\mathbf{x}^-\right)\right)^2\right]\\
&+2\phi_u\phi_l\sum_{i\in\mathcal{Y}}\mathbb{E}_{\substack{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i},\bar{\mathbf{x}}_u\sim\mathbb{P}_{\mathrm{id}},\\\mathbf{x}\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_l),\mathbf{x}^-\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_u)}}\left[\left(\mathbf{h_w}(\mathbf{x})^\top f\left(\mathbf{x}^-\right)\right)^2\right]\\
&+\phi_u^2\mathbb{E}_{\substack{\bar{\mathbf{x}}_u\sim\mathbb{P}_{\mathrm{id}},\bar{\mathbf{x}}_u'\sim\mathbb{P}_{\mathrm{id}},\\\mathbf{x}\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_u),\mathbf{x}^-\sim\mathcal{T}(\cdot|\bar{\mathbf{x}}_u')}}\left[\left(\mathbf{h_w}(\mathbf{x})^\top f\left(\mathbf{x}^-\right)\right)^2\right]\\
=&\phi_l^2\mathcal{L}_3(\mathbf{h_w})+2\phi_u\phi_l\mathcal{L}_4(\mathbf{h_w})+\phi_u^2\mathcal{L}_5(\mathbf{h_w}).
\end{aligned}
$$

We complete this proof. $\qquad\square$

D.2. Derivation of the Data Representations

For the ID representations, we have the following equation.

$$\mathbf{Z}^{(l)} = [\mathbf{D}^{(l)}]^{-\frac{1}{2}} \mathbf{V}_k^{(l)} [\mathbf{\Sigma}_k^{(l)}]^{\frac{1}{2}}. \tag{49}$$

It is easy to see that the following equation holds because of the Eckart–Young–Mirsky theorem (Eckart & Young, 1936) as we use the top-k SVD components of the matrix $\mathbf{A}^{(l)}$ to approximate it (low-rank approximation).

For OOD representations, we have the following equation.

$$\mathbf{Z}_{\text{ood}}^{(l)} = [\mathbf{D}_{\text{ood}}^{(l)}]^{-\frac{1}{2}} \tilde{\mathbf{A}}_{\text{OI}}^{(l)} \mathbf{V}_k^{(l)} [\mathbf{\Sigma}_k^{(l)}]^{-\frac{1}{2}}. \tag{50}$$

The above result can be obtained by the condition that the optimization problem in Equation 7 is solved perfectly, i.e., $\tilde{\mathbf{A}}_{\text{OI}}^{(l)} = \mathbf{F}_{\text{ood}}^{(l)} \mathbf{F}^{(l)\top}$. Note that it is easy to check that $\mathbf{F}^{(l)} = \mathbf{V}_k^{(l)} [\mathbf{\Sigma}_k^{(l)}]^{\frac{1}{2}}$, if we multiply $\mathbf{V}_k^{(l)} [\mathbf{\Sigma}_k^{(l)}]^{-\frac{1}{2}}$ to both sides of the equation $\tilde{\mathbf{A}}_{\text{OI}}^{(l)} = \mathbf{F}_{\text{ood}}^{(l)} \mathbf{F}^{(l)\top}$, we can get:

$$\tilde{\mathbf{A}}_{\text{OI}}^{(l)} \mathbf{V}_k^{(l)} [\mathbf{\Sigma}_k^{(l)}]^{-\frac{1}{2}} = \mathbf{F}_{\text{ood}}^{(l)} [\mathbf{\Sigma}_k^{(l)}]^{\frac{1}{2}} \mathbf{V}_k^{(l)\top} \mathbf{V}_k^{(l)} [\mathbf{\Sigma}_k^{(l)}]^{-\frac{1}{2}} = \mathbf{F}_{\text{ood}}^{(l)}. \tag{51}$$

Note that $\mathbf{F}_{\text{ood}}^{(l)} = [\mathbf{D}_{\text{ood}}^{(l)}]^{\frac{1}{2}} \mathbf{Z}_{\text{ood}}^{(l)}$ and thus we can get $\mathbf{Z}_{\text{ood}}^{(l)} = [\mathbf{D}_{\text{ood}}^{(l)}]^{-\frac{1}{2}} \tilde{\mathbf{A}}_{\text{OI}}^{(l)} \mathbf{V}_k^{(l)} [\mathbf{\Sigma}_k^{(l)}]^{-\frac{1}{2}}$. Therefore we have completed the proof.

Design Rationale. Here we explain the design rationale of the optimization problem in Equation 7 to get the OOD representation $\mathbf{F}_{\text{ood}}^{(l)}$ and $\mathbf{Z}_{\text{ood}}^{(l)}$, which is inspired by the literature in out-of-sample extension.

The out-of-sample extension is a statistical approach that computes the embeddings of new vertices in a graph with the existing in-sample embeddings and the similarity measurements (Bengio et al., 2003; Levin et al., 2018). The goal is to avoid the repeated computational cost of embedding calculations on a large graph when a new vertex emerges. Most of the current works focused on graph Laplacian embedding (Belkin & Niyogi, 2003; Trosset & Priebe, 2008; Belkin et al., 2006; Quispe et al., 2016; Jansen et al., 2015), while a few works relied on the adjacency spectral embedding (Sussman et al., 2012; Tang et al., 2013) for embedding extension. Our framework is similar to the least-squares optimization approach in out-of-sample extension, which derives the out-of-sample embeddings using the adjacency matrix between the in-sample and out-of-sample data (Please refer to Section 3.1 of (Levin et al., 2018) for detailed derivation).

D.3. Upper Bound of the Linear Probing Loss

Lemma 2 (Recap of Lemma 2 in the main paper). *Denote the* $\mathbf{y} \in \mathbb{R}^{(N+M) \times 2}$ *as a matrix contains the one-hot labels for ID and OOD features in* $\mathbf{Z}_{\text{all}}$. *We have:*

$$R(\mathbf{Z}_{\text{all}}) \leq \frac{2}{N+M} \operatorname{Tr} \left(\left(\mathbf{I} - \mathbf{Z}_{\text{all}} \mathbf{Z}_{\text{all}}^{\dagger} \right) \mathbf{y} \mathbf{y}^{\top} \right). \tag{52}$$

Proof. Recall the definition of the linear probing loss is defined as follows:

$$R(\mathbf{Z}_{\text{all}}) \triangleq \min_{\boldsymbol{\theta} \in \mathbb{R}^{k \times 2}} \mathbb{E}_{\mathbf{z} \in \mathbf{z}_{\text{all}}} \mathbb{1}[y(\mathbf{z}) \neq \mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})]. \tag{53}$$

According to Lemma 5.1 in (Sun et al., 2023b), we can get the upper bound of $R(\mathbf{Z}_{\text{all}})$ as follows:

$$R(\mathbf{Z}_{\text{all}}) \leq \frac{2}{N+M} \min_{\boldsymbol{\theta} \in \mathbb{R}^{k \times 2}} \|\mathbf{y} - \mathbf{Z}_{\text{all}} \boldsymbol{\theta}\|_F^2. \tag{54}$$

Given the fact that the closed-form solution of the above minimization problem is $\mathbf{Z}_{\text{all}}^{\dagger} \mathbf{y}$ where $\mathbf{Z}_{\text{all}}^{\dagger}$ denotes the Moore-Penrose inverse of the feature matrix $\mathbf{Z}_{\text{all}}$, we can rewrite the right side of the above inequality as follows:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{k \times 2}} \|\mathbf{y} - \mathbf{Z}_{\text{all}} \boldsymbol{\theta}\|_F^2 = \|\mathbf{y} - \mathbf{Z}_{\text{all}} \mathbf{Z}_{\text{all}}^{\dagger} \mathbf{y}\|_F^2. \tag{55}$$

Simplify it further, we have

$$\|\mathbf{y} - \mathbf{Z}_{\text{all}} \mathbf{Z}_{\text{all}}^{\dagger} \mathbf{y}\|_F^2 = \|(\mathbf{I} - \mathbf{Z}_{\text{all}} \mathbf{Z}_{\text{all}}^{\dagger}) \mathbf{y}\|_F^2 = \operatorname{Tr} \left(\left(\mathbf{I} - \mathbf{Z}_{\text{all}} \mathbf{Z}_{\text{all}}^{\dagger} \right) \mathbf{y} \mathbf{y}^{\top} \right). \tag{56}$$

Therefore we finished the proof. $\qquad \square$

D.4. Necessary Lemmas for Theorem 1

Lemma 3. *Following the definitions in Equations 10 and 28, the error difference for linear probing between the unlabeled and labeled case (with a perturbation magnitude ϕ_l) is upper bounded by the following term:*

$$\mathcal{G}(\phi_l) \geq \frac{2}{3(N+M)} \phi_l \cdot \max \lambda \left(\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l) \right) \cdot \text{Tr} \left(\left(\mathbf{Z}_{\text{all}}(\phi_l) \mathbf{Z}_{\text{all}}(\phi_l)^\top \right)' \mathbf{y}\mathbf{y}^\top \right) \Bigg|_{\phi_l=0}, \tag{57}$$

where $\max \lambda \left(\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l) \right)$ means the maximum eigenvalue of the matrix $\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l)$.

Proof of Lemma 3. From the definition of the derivative, we know that $\mathcal{G}(\phi_l) = - \frac{\mathrm{d}\overline{R}(\mathbf{Z}_{\text{all}}(\phi_l))}{\mathrm{d}\phi_l}\Big|_{\phi_l=0} \cdot \phi_l$, we then investigate the key component $- \frac{\mathrm{d}\overline{R}(\mathbf{Z}_{\text{all}}(\phi_l))}{\mathrm{d}\phi_l}\Big|_{\phi_l=0}$ as follows. Recall that we have

$$-\frac{\mathrm{d}\overline{R}(\mathbf{Z}_{\text{all}}(\phi_l))}{\mathrm{d}\phi_l}\Bigg|_{\phi_l=0} = \text{Tr}\left(\frac{\mathrm{d}\mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger}{\mathrm{d}\phi_l}\mathbf{y}\mathbf{y}^\top \right)\Bigg|_{\phi_l=0} \cdot \frac{2\phi_l}{N+M}$$

$$= \frac{2\phi_l}{N+M} \cdot \text{Tr}\bigg(\Big[\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger + \mathbf{Z}_{\text{all}}(\phi_l)\Big(-\big(\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l)\big)^{-1}\big(\mathbf{Z}'_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l)$$

$$+ \mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}'_{\text{all}}(\phi_l)\big)\big(\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l)\big)^{-1}\mathbf{Z}_{\text{all}}(\phi_l)^\top + \big(\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l)\big)^{-1}\mathbf{Z}'_{\text{all}}(\phi_l)^\top\Big)\Big]\mathbf{y}\mathbf{y}^\top \bigg)\Bigg|_{\phi_l=0}$$

$$= \frac{2\phi_l}{N+M}\text{Tr}\bigg(\Big[\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger - \big(\mathbf{Z}_{\text{all}}(\phi_l)^\dagger\big)^\top \mathbf{Z}'_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger$$

$$- \big(\mathbf{Z}_{\text{all}}(\phi_l)^\dagger\big)^\top \mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger + \big(\mathbf{Z}_{\text{all}}(\phi_l)^\dagger\big)^\top \mathbf{Z}'_{\text{all}}(\phi_l)^\top\Big]\mathbf{y}\mathbf{y}^\top \bigg)\Bigg|_{\phi_l=0}$$

$$= \frac{2\phi_l}{N+M}\text{Tr}\bigg(\Big[\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger - \big(\mathbf{Z}_{\text{all}}(\phi_l)^\dagger\big)^\top \mathbf{Z}'_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger$$

$$- (\mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)^\top \mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger + \big(\mathbf{Z}_{\text{all}}(\phi_l)^\dagger\big)^\top \mathbf{Z}'_{\text{all}}(\phi_l)^\top\Big]\mathbf{y}\mathbf{y}^\top \bigg)\Bigg|_{\phi_l=0}$$

$$= \frac{2\phi_l}{N+M}\text{Tr}\left(\Big[(\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)^\top \mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger + \big(\mathbf{Z}_{\text{all}}(\phi_l)^\dagger\big)^\top \mathbf{Z}'_{\text{all}}(\phi_l)^\top(\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)\Big]\mathbf{y}\mathbf{y}^\top \right)\Bigg|_{\phi_1=0}$$

$$= \frac{2\phi_l}{N+M}\text{Tr}\left(\Big[(\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger + \big(\mathbf{Z}_{\text{all}}(\phi_l)^\dagger\big)^\top \mathbf{Z}'_{\text{all}}(\phi_l)^\top(\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)^\top\Big]\mathbf{y}\mathbf{y}^\top \right)\Bigg|_{\phi_1=0} \tag{58}$$

$$= \frac{2\phi_l}{N+M}\text{Tr}\left(\Big[(\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger + \big[(\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger\big]^\top\Big]\mathbf{y}\mathbf{y}^\top \right)\Bigg|_{\phi_1=0}$$

$$= \frac{2\phi_l}{N+M}\text{Tr}\left((\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger) \cdot \Big(\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger + (\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)^\top\Big)\mathbf{y}\mathbf{y}^\top \right)\Bigg|_{\phi_1=0}$$

$$\geq \frac{2\phi_l}{N+M}\text{Tr}\left((\tfrac{2}{3}\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger) \cdot \Big(\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger + (\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)^\top\Big)\mathbf{y}\mathbf{y}^\top \right)\Bigg|_{\phi_1=0}$$

$$\geq \frac{2\phi_l}{N+M}\Big[\nu(\tfrac{2}{3}\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger) + \pi(\tfrac{2}{3}\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)\Big] \cdot$$

$$\text{Tr}\left(\Big(\mathbf{Z}'_{\text{all}}(\phi_l)(\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l))^{-1}\mathbf{Z}_{\text{all}}(\phi_l)^\top + \mathbf{Z}_{\text{all}}(\phi_l)(\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l))^{-1}\mathbf{Z}'_{\text{all}}(\phi_l)^\top\Big)\mathbf{y}\mathbf{y}^\top \right)\Bigg|_{\phi_1=0}$$

$$= \frac{2\phi_l}{N+M}\frac{1}{3} \cdot \text{Tr}\left(\Big((\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l))^{-1}\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}'_{\text{all}}(\phi_l) + (\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l))^{-1}\mathbf{Z}'_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l)\Big)\mathbf{y}\mathbf{y}^\top \right)\Bigg|_{\phi_1=0}$$

$$\geq \frac{2\phi_l}{N+M}\frac{1}{3}\min\lambda((\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l))^{-1}) \cdot \text{Tr}\left(\Big(\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}'_{\text{all}}(\phi_l) + \mathbf{Z}'_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l)\Big)\mathbf{y}\mathbf{y}^\top \right)\Bigg|_{\phi_1=0}$$

$$= \frac{2\phi_l}{3(N+M)}\max\lambda(\mathbf{Z}_{\text{all}}(\phi_l)^\top \mathbf{Z}_{\text{all}}(\phi_l)) \cdot \text{Tr}\left(\Big((\mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\top)'\Big)\mathbf{y}\mathbf{y}^\top \right)\Bigg|_{\phi_1=0}$$

where $\nu(\tfrac{2}{3}\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)$ and $\pi(\tfrac{2}{3}\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^\dagger)$ denote the smallest negative eigenvalues and the smallest

positive eigenvalues[1] of the matrix $\frac{2}{3}\mathbf{I} - \mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^{\dagger}$.

The first seven equations calculate the derivative w.r.t. the perturbation magnitude ϕ_l and use the cyclic property of the trace operator. The first inequality is true because we have

$$\frac{1}{3}\text{Tr}((\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^{\dagger} + (\mathbf{Z}'_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^{\dagger})^{\top})\mathbf{y}\mathbf{y}^{\top})\big|_{\phi_l=0} \geq \frac{1}{3}\max\lambda(\mathbf{Z}_{\text{all}}(\phi_l)^{\top}\mathbf{Z}_{\text{all}}(\phi_l))\max\lambda(\mathbf{y}\mathbf{y}^{\top})\cdot\text{Tr}((\mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^{\top})')\big|_{\phi_l=0}. \quad (59)$$

Since we know $\max\lambda(\mathbf{Z}_{\text{all}}(\phi_l)^{\top}\mathbf{Z}_{\text{all}}(\phi_l))\big|_{\phi_l=0} > 0$ (Lemma 4) and $\max\lambda(\mathbf{y}\mathbf{y}^{\top}) > 0$, and $\text{Tr}((\mathbf{Z}_{\text{all}}(\phi_l)\mathbf{Z}_{\text{all}}(\phi_l)^{\top})')\big|_{\phi_l=0}$ is lower bounded and can be large than 0 (Lemmas 5 and 6). The second inequality holds because of the main theorem in (Baksalary & Puntanen, 1992). The third inequality holds because the inequality for the trace of matrix product of two square matrices (Fang et al., 1994), i.e., $\text{Tr}(\mathbf{H}\mathbf{M}) \geq \min\lambda(\mathbf{H})\cdot\text{Tr}(\mathbf{M})$. $\qquad\square$

Lemma 4. *Suppose the perturbation magnitude ϕ_l is 0, denote the input feature of linear probing as $\mathbf{Z}_{\text{all}}(0)$, then we have that:*

$$\max\lambda\left(\mathbf{Z}_{\text{all}}(\phi_l)^{\top}\mathbf{Z}_{\text{all}}(\phi_l)\right)\bigg|_{\phi_l=0} > 0, \quad (60)$$

here $\max\lambda\left(\mathbf{Z}_{\text{all}}(\phi_l)^{\top}\mathbf{Z}_{\text{all}}(\phi_l)\right)$ means the maximum eigenvalue of the matrix $\mathbf{Z}_{\text{all}}(\phi_l)^{\top}\mathbf{Z}_{\text{all}}(\phi_l)$.

Proof of Lemma 4. Since we know that the column rank of the ID feature matrix $\mathbf{Z}(0) \in \mathbb{R}^{N\times k}$ is k, then when adding M rows of OOD features $\mathbf{Z}_{\text{ood}}(0)$ to get $\mathbf{Z}_{\text{all}}(0)$ will not increase the column rank. Therefore, the rank of matrix $\mathbf{Z}_{\text{all}}(0)$ is equal to k, which means $\mathbf{Z}_{\text{all}}(0)$ has a full column rank. Thus, the matrix $\mathbf{Z}_{\text{all}}(0)^{\top}\mathbf{Z}_{\text{all}}(0)$ is positive definite [2], which means the eigenvalues of $\mathbf{Z}_{\text{all}}(0)^{\top}\mathbf{Z}_{\text{all}}(0)$ are all greater than 0 and the lemma is proved. $\qquad\square$

Lemma 5. *If Assumption 2 holds, the lower bound for $\text{Tr}\left(\left(\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^{\top}\right)'\right)\bigg|_{\phi_l=0}$ is given as follows:*

$$\text{Tr}\left(\left(\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^{\top}\right)'\right)\bigg|_{\phi_l=0} \geq \sum_{i=1}^{c}\|\mathfrak{q}_i\|_F^2 - 2r^2\sum_{i=1}^{c}\|\mathfrak{q}_i\|_1, \quad (61)$$

where $\mathfrak{q} \in \mathbb{R}^{N\times c}$ with each column defined as $(\mathfrak{q}_i)_{\mathbf{x}} = \mathbb{E}_{\bar{\mathbf{x}}_l\sim\mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l), \mathbf{x} \in \mathcal{X}_{\text{id}}$, and r is the maximum l_2 norm of the ID representations, i.e., $r = \max_{\mathbf{z}\in\mathbf{Z}^{(u)}}\|\mathbf{z}\|_2$.

Proof of Lemma 5. As $\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^{\top}$ denotes the top-k SVD components of the matrix $\tilde{\mathbf{A}}(\phi_l)$, according to Equation 26, we rewrite it using the eigenvalues and the eigen projectors as follows.

$$\begin{aligned}\text{Tr}\left(\left(\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^{\top}\right)'\right)\bigg|_{\phi_l=0} &= \sum_{j=1}^{k}\text{Tr}\left([\mathbf{D}(\phi_l)^{-\frac{1}{2}}\boldsymbol{\lambda}_j(\phi_l)\boldsymbol{\Phi}_j(\phi_l)\mathbf{D}(\phi_l)^{-\frac{1}{2}}]'\right)\bigg|_{\phi_l=0} \\ &= \sum_{j=1}^{k}\text{Tr}\left(\left[\mathbf{D}(\phi_l)^{-\frac{1}{2}}\right]'\boldsymbol{\lambda}_j^{(u)}\boldsymbol{\Phi}_j^{(u)} + \boldsymbol{\lambda}_j^{(u)}\boldsymbol{\Phi}_j^{(u)}\left[\mathbf{D}(\phi_l)^{-\frac{1}{2}}\right]' + [\boldsymbol{\lambda}_j(\phi_l)]'\boldsymbol{\Phi}_j^{(u)} + \boldsymbol{\lambda}_j^{(u)}[\boldsymbol{\Phi}_j(\phi_l)]'\right)\bigg|_{\phi_l=0},\end{aligned} \quad (62)$$

where $\boldsymbol{\lambda}_j^{(u)}$ and $\boldsymbol{\Phi}_j^{(u)}$ are the j-th eigenvalue and eigen projector of matrix $\tilde{\mathbf{A}}(\phi_l)$ when $\phi_l = 0$. Moreover, assume $\tilde{\mathbf{A}}^{(u)}$ to be the normalized adjacency matrix when $\phi_l = 0$, recall $\mathbf{D}(\phi_l) = \mathbf{I}_{N\times N} + \phi_l\mathbf{D}_l$ where $\mathbf{D}_l = \sum_{i=1}^{c}\text{diag}(\mathfrak{q}_i)$ (Equation 24), we have the following calculation for the derivatives:

$$\left[\mathbf{D}(\phi_l)^{-\frac{1}{2}}\right]' = -\frac{1}{2}\mathbf{D}_l. \quad (63)$$

---

[1] The value of the two terms are $\frac{2}{3}$ and $-\frac{1}{3}$ according to https://math.stackexchange.com/questions/188129

[2] Proof idea is shown here https://math.stackexchange.com/questions/2202242

$$\left.[\tilde{\mathbf{A}}(\phi_l)]'\right|_{\phi_l=0} = \left.\left[\mathbf{D}(\phi_l)^{-\frac{1}{2}}\mathbf{A}(\phi_l)\mathbf{D}(\phi_l)^{-\frac{1}{2}}\right]'\right|_{\phi_l=0}$$

$$= \left[\mathbf{D}(\phi_l)^{-\frac{1}{2}}\right]'\tilde{\mathbf{A}}^{(u)} + [\mathbf{A}(\phi_l)]' + \tilde{\mathbf{A}}^{(u)}\left[\mathbf{D}(\phi_l)^{-\frac{1}{2}}\right]' \quad (64)$$

$$= -\frac{1}{2}\mathbf{D}_l\tilde{\mathbf{A}}^{(u)} + \sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \frac{1}{2}\tilde{\mathbf{A}}^{(u)}\mathbf{D}_l$$

To get the derivative of the eigenvalues, according to Equation 3 in (Greenbaum et al., 2020),

$$\left.[\boldsymbol{\lambda}_j(\phi_l)]'\right|_{\phi_l=0} = \left.\mathrm{Tr}\left(\boldsymbol{\Phi}_j^{(u)}[\tilde{\mathbf{A}}(\phi_l)]'\right)\right|_{\phi_l=0}$$

$$= \mathrm{Tr}\left(\boldsymbol{\Phi}_j^{(u)}\left(-\frac{1}{2}\mathbf{D}_l\tilde{\mathbf{A}}^{(u)} + \sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \frac{1}{2}\tilde{\mathbf{A}}^{(u)}\mathbf{D}_l\right)\right)$$

$$= \mathrm{Tr}\left(-\frac{\boldsymbol{\lambda}_j^{(u)}}{2}\mathbf{D}_l\boldsymbol{\Phi}_j^{(u)} + \boldsymbol{\Phi}_j^{(u)}\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \frac{\boldsymbol{\lambda}_j^{(u)}}{2}\boldsymbol{\Phi}_j^{(u)}\mathbf{D}_l\right) \quad (65)$$

$$= \mathrm{Tr}\left(\boldsymbol{\Phi}_j^{(u)}\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \boldsymbol{\lambda}_j^{(u)}\mathbf{D}_l\right)\right).$$

To get the derivative of the eigenvectors, according to Equation 10 in (Greenbaum et al., 2020),

$$\left.[\boldsymbol{\Phi}_j(\phi_l)]'\right|_{\phi_l=0} = \left(\boldsymbol{\lambda}_j^{(u)}\mathbf{I}_N - \tilde{\mathbf{A}}^{(u)}\right)^{\dagger}[\tilde{\mathbf{A}}(\phi_l)]'\boldsymbol{\Phi}_j^{(u)} + \boldsymbol{\Phi}_j^{(u)}[\tilde{\mathbf{A}}(\phi_l)]'\left(\boldsymbol{\lambda}_j^{(u)}\mathbf{I}_N - \tilde{\mathbf{A}}^{(u)}\right)^{\dagger}$$

$$= \sum_{i\neq j}^{N}\frac{1}{\boldsymbol{\lambda}_j^{(u)} - \boldsymbol{\lambda}_i^{(u)}}\left(\boldsymbol{\Phi}_i^{(u)}[\tilde{\mathbf{A}}(\phi_l)]'\boldsymbol{\Phi}_j^{(u)} + \boldsymbol{\Phi}_j^{(u)}[\tilde{\mathbf{A}}(\phi_l)]'\boldsymbol{\Phi}_i^{(u)}\right)$$

$$= \sum_{i\neq j}^{N}\frac{1}{\boldsymbol{\lambda}_j^{(u)} - \boldsymbol{\lambda}_i^{(u)}}\left(\boldsymbol{\Phi}_i^{(u)}\left(-\frac{1}{2}\mathbf{D}_l\tilde{\mathbf{A}}^{(u)} + \sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \frac{1}{2}\tilde{\mathbf{A}}^{(u)}\mathbf{D}_l\right)\boldsymbol{\Phi}_j^{(u)} + \boldsymbol{\Phi}_j^{(u)}(\ldots)\boldsymbol{\Phi}_i^{(u)}\right)$$

$$= \sum_{i\neq j}^{N}\frac{1}{\boldsymbol{\lambda}_j^{(u)} - \boldsymbol{\lambda}_i^{(u)}}\left(\boldsymbol{\Phi}_i^{(u)}\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \frac{\boldsymbol{\lambda}_j^{(u)} + \boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\boldsymbol{\Phi}_j^{(u)} + \boldsymbol{\Phi}_j^{(u)}\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \frac{\boldsymbol{\lambda}_j^{(u)} + \boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\boldsymbol{\Phi}_i^{(u)}\right).$$

$$(66)$$

Put them together in Equation 62, we can get

$$\left.\mathrm{Tr}\left(\left(\mathbf{V}_k(\phi_l)\boldsymbol{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^{\top}\right)'\right)\right|_{\phi_l=0}$$

$$= \left.\sum_{j=1}^{k}\mathrm{Tr}\left(-\frac{1}{2}\mathbf{D}_l\boldsymbol{\lambda}_j^{(u)}\boldsymbol{\Phi}_j^{(u)} - \frac{1}{2}\boldsymbol{\lambda}_j^{(u)}\boldsymbol{\Phi}_j^{(u)}\mathbf{D}_l + \mathrm{Tr}\left(\boldsymbol{\Phi}_j^{(u)}\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \boldsymbol{\lambda}_j^{(u)}\mathbf{D}_l\right)\right)\boldsymbol{\Phi}_j^{(u)} + \boldsymbol{\lambda}_j^{(u)}[\boldsymbol{\Phi}_j(\phi_l)]'\right)\right|_{\phi_l=0}$$

$$= \left.\left[\sum_{j=1}^{k}-\frac{\boldsymbol{\lambda}_j^{(u)}}{2}\mathrm{Tr}\left(\mathbf{D}_l\boldsymbol{\Phi}_j^{(u)} + \boldsymbol{\Phi}_j^{(u)}\mathbf{D}_l\right) + \mathrm{Tr}\left(\boldsymbol{\Phi}_j^{(u)}\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \boldsymbol{\lambda}_j^{(u)}\mathbf{D}_l\right)\right)\mathrm{Tr}\left(\boldsymbol{\Phi}_j^{(u)}\right) + \mathrm{Tr}\left(\boldsymbol{\lambda}_j^{(u)}[\boldsymbol{\Phi}_j(\phi_l)]'\right)\right]\right|_{\phi_l=0}$$

$$(67)$$

For $\mathrm{Tr}\left(\boldsymbol{\Phi}_j^{(u)}\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \boldsymbol{\lambda}_j^{(u)}\mathbf{D}_l\right)\right)\mathrm{Tr}\left(\boldsymbol{\Phi}_j^{(u)}\right)$, we have the following:

$$\sum_{j=1}^{k}\mathrm{Tr}\left(\boldsymbol{\Phi}_j^{(u)}\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \boldsymbol{\lambda}_j^{(u)}\mathbf{D}_l\right)\right)\mathrm{Tr}\left(\boldsymbol{\Phi}_j^{(u)}\right) = \sum_{j=1}^{k}\mathrm{Tr}\left(\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top} - \boldsymbol{\lambda}_j^{(u)}\mathbf{D}_l\right)\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_j^{(u)}\right) \quad (68)$$

For $\mathrm{Tr}\left(\boldsymbol{\lambda}_j^{(u)}\left[\boldsymbol{\Phi}_j(\phi_l)\right]'\right)$, we have the following:

$$
\begin{aligned}
&\sum_{j=1}^k \mathrm{Tr}\left(\boldsymbol{\lambda}_j^{(u)}\left[\boldsymbol{\Phi}_j(\phi_l)\right]'\right)\\
&=\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i\neq j}^N \frac{\boldsymbol{\lambda}_j^{(u)}}{\boldsymbol{\lambda}_j^{(u)}-\boldsymbol{\lambda}_i^{(u)}}\left(\boldsymbol{\Phi}_i^{(u)}\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\boldsymbol{\Phi}_j^{(u)}+\boldsymbol{\Phi}_j^{(u)}\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\boldsymbol{\Phi}_i^{(u)}\right)\right)\\
&=\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i\neq j}^N \frac{\boldsymbol{\lambda}_j^{(u)}}{\boldsymbol{\lambda}_j^{(u)}-\boldsymbol{\lambda}_i^{(u)}}\left(\left(\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_i^{(u)}+\boldsymbol{\Phi}_i^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right)\\
&=\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i\neq j,i\leq k} \frac{\boldsymbol{\lambda}_j^{(u)}}{\boldsymbol{\lambda}_j^{(u)}-\boldsymbol{\lambda}_i^{(u)}}\left(\left(\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_i^{(u)}+\boldsymbol{\Phi}_i^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right)\\
&\quad+\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i=k+1}^N \frac{\boldsymbol{\lambda}_j^{(u)}}{\boldsymbol{\lambda}_j^{(u)}-\boldsymbol{\lambda}_i^{(u)}}\left(\left(\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_i^{(u)}+\boldsymbol{\Phi}_i^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right)\\
&=\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i<j}\left(\frac{\boldsymbol{\lambda}_j^{(u)}}{\boldsymbol{\lambda}_j^{(u)}-\boldsymbol{\lambda}_i^{(u)}}+\frac{\boldsymbol{\lambda}_i^{(u)}}{\boldsymbol{\lambda}_i^{(u)}-\boldsymbol{\lambda}_j^{(u)}}\right)\left(\left(\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_i^{(u)}+\boldsymbol{\Phi}_i^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right)\\
&\quad+\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i=k+1}^N \frac{\boldsymbol{\lambda}_j^{(u)}}{\boldsymbol{\lambda}_j^{(u)}-\boldsymbol{\lambda}_i^{(u)}}\left(\left(\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_i^{(u)}+\boldsymbol{\Phi}_i^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right)\\
&=\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i<j}\left(\left(\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_i^{(u)}+\boldsymbol{\Phi}_i^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right)\\
&\quad+\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i=k+1}^N \frac{\boldsymbol{\lambda}_j^{(u)}}{\boldsymbol{\lambda}_j^{(u)}-\boldsymbol{\lambda}_i^{(u)}}\left(\left(\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_i^{(u)}+\boldsymbol{\Phi}_i^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right)\\
&=\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i\neq j,i\leq k}\frac{1}{2}\left(\left(\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_i^{(u)}+\boldsymbol{\Phi}_i^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right)\\
&\quad+\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i=k+1}^N \frac{\boldsymbol{\lambda}_j^{(u)}}{\boldsymbol{\lambda}_j^{(u)}-\boldsymbol{\lambda}_i^{(u)}}\left(\left(\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_i^{(u)}+\boldsymbol{\Phi}_i^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right)
\end{aligned}
\tag{69}
$$

If we put the result in Equations 68 and 69 together, we can get:

$$
\begin{aligned}
&\sum_{j=1}^k \mathrm{Tr}\left(\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\boldsymbol{\lambda}_j^{(u)}\mathbf{D}_l\right)\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)+\sum_{j=1}^k \mathrm{Tr}\left(\boldsymbol{\lambda}_j^{(u)}\left[\boldsymbol{\Phi}_j(\phi_l)\right]'\right)\\
&=\sum_{j=1}^k \mathrm{Tr}\left(\frac{\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_j^{(u)}+\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_j^{(u)}}{2}\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_j^{(u)}}{2}\mathbf{D}_l\right)\right)+\sum_{j=1}^k \mathrm{Tr}\left(\boldsymbol{\lambda}_j^{(u)}\left[\boldsymbol{\Phi}_j(\phi_l)\right]'\right)\\
&=\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i\leq k}\frac{1}{2}\left(\left(\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_i^{(u)}+\boldsymbol{\Phi}_i^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right)\\
&\quad+\sum_{j=1}^k \mathrm{Tr}\left(\sum_{i=k+1}^N \frac{\boldsymbol{\lambda}_j^{(u)}}{\boldsymbol{\lambda}_j^{(u)}-\boldsymbol{\lambda}_i^{(u)}}\left(\left(\boldsymbol{\Phi}_j^{(u)}\boldsymbol{\Phi}_i^{(u)}+\boldsymbol{\Phi}_i^{(u)}\boldsymbol{\Phi}_j^{(u)}\right)\left(\sum_{i=1}^c \mathsf{q_iq_i}^\top-\frac{\boldsymbol{\lambda}_j^{(u)}+\boldsymbol{\lambda}_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right)
\end{aligned}
\tag{70}
$$

For the term $-\frac{\lambda_j^{(u)}}{2}\sum_{j=1}^{k}\mathrm{Tr}\left(\mathbf{D}_l\mathbf{\Phi}_j^{(u)}+\mathbf{\Phi}_j^{(u)}\mathbf{D}_l\right)$, we can have the following equation:

$$
\begin{aligned}
-\frac{\lambda_j^{(u)}}{2}\sum_{j=1}^{k}\mathrm{Tr}\left(\mathbf{D}_l\mathbf{\Phi}_j^{(u)}+\mathbf{\Phi}_j^{(u)}\mathbf{D}_l\right) &= -\sum_{j=1}^{k}\frac{\lambda_j^{(u)}}{2}\mathrm{Tr}\left(\mathbf{D}_l\mathbf{\Phi}_j^{(u)}+\mathbf{\Phi}_j^{(u)}\mathbf{D}_l\right) \\
&= -\sum_{j=1}^{k}\frac{\lambda_j^{(u)}}{2}\mathrm{Tr}\left((\mathbf{\Phi}_j^{(u)}+\mathbf{\Phi}_j^{(u)})\mathbf{D}_l\right) \\
&= -\sum_{j=1}^{k}\frac{\lambda_j^{(u)}}{2}\mathrm{Tr}\left((\mathbf{\Phi}_j^{(u)}\mathbf{I}+\mathbf{I}\mathbf{\Phi}_j^{(u)})\mathbf{D}_l\right) \\
&= -\sum_{j=1}^{k}\frac{\lambda_j^{(u)}}{2}\mathrm{Tr}\left((\mathbf{\Phi}_j^{(u)}\sum_{i=1}^{N}\mathbf{\Phi}_i^{(u)}+\sum_{i=1}^{N}\mathbf{\Phi}_i^{(u)}\mathbf{\Phi}_j^{(u)})\mathbf{D}_l\right) \\
&= -\sum_{j=1}^{k}\mathrm{Tr}\left(\sum_{i=1}^{N}\frac{\lambda_j^{(u)}}{2}(\mathbf{\Phi}_j^{(u)}\mathbf{\Phi}_i^{(u)}+\mathbf{\Phi}_i^{(u)}\mathbf{\Phi}_j^{(u)})\mathbf{D}_l\right)
\end{aligned}
\tag{71}
$$

Therefore, we can get the following result for $\mathrm{Tr}\left(\left(\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^{\top}\right)'\right)\Big|_{\phi_l=0}$:

$$
\begin{aligned}
\mathrm{Tr}\left(\left(\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^{\top}\right)'\right)\Big|_{\phi_l=0} &= \left[\sum_{j=1}^{k}\mathrm{Tr}\left(\sum_{i=1}^{k}\frac{1}{2}\left((\mathbf{\Phi}_j^{(u)}\mathbf{\Phi}_i^{(u)}+\mathbf{\Phi}_i^{(u)}\mathbf{\Phi}_j^{(u)})\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top}-\frac{3\lambda_j^{(u)}+\lambda_i^{(u)}}{2}\mathbf{D}_l\right)\right)\right) \right. \\
&\quad \left. +\sum_{j=1}^{k}\mathrm{Tr}\left(\sum_{i=k+1}^{N}\frac{\lambda_j^{(u)}}{\lambda_j^{(u)}-\lambda_i^{(u)}}\left((\mathbf{\Phi}_j^{(u)}\mathbf{\Phi}_i^{(u)}+\mathbf{\Phi}_i^{(u)}\mathbf{\Phi}_j^{(u)})\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top}-\lambda_j^{(u)}\mathbf{D}_l\right)\right)\right)\right] \\
&= \left[\sum_{j=1}^{k}\mathrm{Tr}\left(\sum_{i=1}^{k}\frac{1}{2}\left((\mathbf{\Phi}_j^{(u)}\mathbf{\Phi}_i^{(u)}+\mathbf{\Phi}_i^{(u)}\mathbf{\Phi}_j^{(u)})\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top}-2\lambda_j^{(u)}\mathbf{D}_l\right)\right)\right)\right. \\
&\quad \left. +\sum_{j=1}^{k}\mathrm{Tr}\left(\sum_{i=k+1}^{N}\frac{\lambda_j^{(u)}}{\lambda_j^{(u)}-\lambda_i^{(u)}}\left((\mathbf{\Phi}_j^{(u)}\mathbf{\Phi}_i^{(u)}+\mathbf{\Phi}_i^{(u)}\mathbf{\Phi}_j^{(u)})\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top}-\lambda_j^{(u)}\mathbf{D}_l\right)\right)\right)\right] \\
&= \left[\sum_{j=1}^{k}\sum_{i=1}^{k}\mathrm{Tr}\left(\mathsf{v}_j^{(u)\top}\mathsf{v}_i^{(u)}\cdot\mathsf{v}_i^{(u)\top}\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top}-2\lambda_j^{(u)}\mathbf{D}_l\right)\mathsf{v}_j^{(u)}\right)\right. \\
&\quad \left. +\sum_{j=1}^{k}\sum_{i=k+1}^{N}\mathrm{Tr}\left(\frac{2\lambda_j^{(u)}}{\lambda_j^{(u)}-\lambda_i^{(u)}}\mathsf{v}_j^{(u)\top}\mathsf{v}_i^{(u)}\cdot\mathsf{v}_i^{(u)\top}\left(\sum_{i=1}^{c}\mathsf{q}_i\mathsf{q}_i^{\top}-\lambda_j^{(u)}\mathbf{D}_l\right)\mathsf{v}_j^{(u)}\right)\right].
\end{aligned}
\tag{72}
$$

Since $\tilde{\mathbf{A}}^{(u)}$ has all positive eigenvalues, then we can get $\frac{\lambda_j^{(u)}}{\lambda_j^{(u)}-\lambda_i^{(u)}}\geq 1$. Denote $\mathbf{V}_k^{(u)}$ and $\mathbf{\Sigma}_k^{(u)}$ as the matrix of the first k eigenvectors and eigenvalues of $\tilde{\mathbf{A}}^{(u)}$ when $\phi_l=0$, if we rewrite the above formula in the matrix form, we can get the

following:

$$\text{Tr}\left(\left(\mathbf{V}_k(\phi_l)\boldsymbol{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^\top\right)'\right)\bigg|_{\phi_l=0} \geq \left[\text{Tr}(\mathbf{V}_k^{(u)\top}\mathbf{V}_k^{(u)} \cdot \mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top\mathbf{V}_k^{(u)}) - 2\,\text{Tr}(\mathbf{V}_k^{(u)\top}\mathbf{V}_k^{(u)} \cdot \boldsymbol{\Sigma}_k^{(u)}\mathbf{V}_k^{(u)\top}\mathbf{D}_l\mathbf{V}_k^{(u)})\right.$$

$$+ 2\,\text{Tr}(\mathbf{V}_\varnothing^{(u)\top}\mathbf{V}_k^{(u)} \cdot \mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top\mathbf{V}_\varnothing^{(u)}) - 2\,\text{Tr}(\mathbf{V}_\varnothing^{(u)\top}\mathbf{V}_k^{(u)} \cdot \boldsymbol{\Sigma}_k^{(u)}\mathbf{V}_k^{(u)\top}\mathbf{D}_l\mathbf{V}_\varnothing^{(u)})$$

$$= \left[\text{Tr}(\mathbf{V}_k^{(u)} \cdot \mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}) - 2\,\text{Tr}(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\mathbf{V}_k^{(u)} \cdot \mathbf{V}_k^{(u)\top})\right.$$

$$+ 2\,\text{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top\left(\mathbf{I}_N - \mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\right)\right) - 2\,\text{Tr}\left(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\left(\mathbf{I}_N - \mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\right)\right)\right]$$

$$= \left[2\,\text{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top\right) - 2\,\text{Tr}\left(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\right) - \text{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\right)\right]$$

$$= \left[2\,\text{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top - \tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l - \frac{1}{2}\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\right)\right]$$

$$= \left[\text{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top - 2\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l + \mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top\mathbf{V}_\varnothing^{(u)}\mathbf{V}_\varnothing^{(u)\top}\right)\right].$$

$$(73)$$

For $\text{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top - 2\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l + \mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top\mathbf{V}_\varnothing^{(u)}\mathbf{V}_\varnothing^{(u)\top}\right)$, given that we let $\mathbf{D}_l \triangleq \sum_{i=1}^c \text{diag}(\mathfrak{q}_i)$, and Assumption 2, i.e., $\mathfrak{q}_i^\top\mathbf{V}_\varnothing^{(u)}=0$, we have that:

$$\text{Tr}\left(\left(\mathbf{V}_k(\phi_l)\boldsymbol{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^\top\right)'\right)\bigg|_{\phi_l=0} \geq \text{Tr}\left(\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top - 2\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\right) = \text{Tr}\left(\sum_{i=1}^c \mathfrak{q}_i\mathfrak{q}_i^\top - 2\mathbf{F}^{(u)}\mathbf{F}^{(u)\top}\mathbf{D}_l\right)$$

$$= \sum_{i=1}^c \|\mathfrak{q}_i\|_F^2 - 2\mathbf{F}_i^{(u)\top}\mathbf{F}_i^{(u)}\sum_{i=1}^c \|\mathfrak{q}_i\|_1,$$

$$(74)$$

According to the definition of r, we have the upper bound of the ID features as $\mathbf{F}_i^{(u)\top}\mathbf{F}_i^{(u)} \leq r^2$. Therefore, we have the following lower bound:

$$\text{Tr}\left(\left(\mathbf{V}_k(\phi_l)\boldsymbol{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^\top\right)'\right)\bigg|_{\phi_l=0} \geq \sum_{i=1}^c \|\mathfrak{q}_i\|_F^2 - 2r^2\sum_{i=1}^c \|\mathfrak{q}_i\|_1. \tag{75}$$

$$\square$$

Lemma 6. *Under same conditions in Lemma 5, the lower bound for* $\text{Tr}\left(\left(\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\boldsymbol{\Sigma}_k(\phi_l)^{-1}\mathbf{V}_k(\phi_l)^\top\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)^\top\right)'\right)\bigg|_{\phi_l=0}$ *is given as follows:*

$$\text{Tr}\left(\left(\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\boldsymbol{\Sigma}_k(\phi_l)^{-1}\mathbf{V}_k(\phi_l)^\top\tilde{\mathbf{A}}_{\text{OI}}(\phi_l)^\top\right)'\right)\bigg|_{\phi_l=0}$$

$$\geq 2\sum_{i=1}^c \text{Tr}\left(\mathfrak{p}_i\mathfrak{q}_i^\top \cdot \tilde{\mathbf{A}}_{\text{OI}}^{(u)\top}\right) - \|\tilde{\mathbf{A}}_{\text{OI}}^{(u)}\|_F^2\|\tilde{\mathbf{A}}^{(u)}\|_F^2 \cdot \left(\sum_{i=1}^c \|\mathfrak{q}_i\|_F^2 - \frac{2r^2(\tau-k)}{\tau-1}\sum_{i=1}^c \|\mathfrak{q}_i\|_1\right),$$

$$(76)$$

where $\mathfrak{p} \in \mathbb{R}^{M \times c}$ *is defined as* $(\mathfrak{p}_i)_{\mathbf{x}} = \mathbb{E}_{\bar{\mathbf{x}}_l \sim \mathbb{P}_{l_i}}\mathcal{T}(\mathbf{x}|\bar{\mathbf{x}}_l), \mathbf{x} \in \mathcal{X}_{\text{ood}}.$

Proof of Lemma 6. Given the fact that $\tilde{\mathbf{A}}_{\mathrm{OI}}(\phi_l) = \phi_u \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)} + \phi_l \cdot \sum_{i=1}^{c} \mathfrak{p}_i \mathfrak{q}_i^\top$, we have the following:

$$
\begin{aligned}
&\mathrm{Tr}\left(\left(\tilde{\mathbf{A}}_{\mathrm{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)^{-1}\mathbf{V}_k(\phi_l)^\top \tilde{\mathbf{A}}_{\mathrm{OI}}(\phi_l)^\top\right)'\right)\bigg|_{\phi_l=0} \\
&= 2\,\mathrm{Tr}\left(\sum_{i=1}^{c}\mathfrak{p}_i\mathfrak{q}_i^\top \cdot \mathbf{V}_k^{(u)}\mathbf{\Sigma}_k^{(u)-1}\mathbf{V}_k^{(u)\top}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) + \mathrm{Tr}\left(\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\left(\mathbf{V}_k(\phi_l)\mathbf{\Sigma}_k(\phi_l)^{-1}\mathbf{V}_k(\phi_l)^\top\right)'\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right)\bigg|_{\phi_l=0} \\
&= 2\,\mathrm{Tr}\left(\sum_{i=1}^{c}\mathfrak{p}_i\mathfrak{q}_i^\top \cdot (\tilde{\mathbf{A}}^{(u)})^{-1}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) + \mathrm{Tr}\left((\tilde{\mathbf{A}}_k(\phi_l)^{-1})' \cdot \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\right)\bigg|_{\phi_l=0} \\
&\geq 2\,\mathrm{Tr}\left(\sum_{i=1}^{c}\mathfrak{p}_i\mathfrak{q}_i^\top \cdot (\tilde{\mathbf{A}}^{(u)})^{-1}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) + \min\lambda(\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}) \cdot \mathrm{Tr}\left((\tilde{\mathbf{A}}_k(\phi_l)^{-1})'\right)\bigg|_{\phi_l=0} \\
&= 2\,\mathrm{Tr}\left(\sum_{i=1}^{c}\mathfrak{p}_i\mathfrak{q}_i^\top \cdot (\tilde{\mathbf{A}}^{(u)})^{-1}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) - \min\lambda(\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}) \cdot \mathrm{Tr}\left((\tilde{\mathbf{A}}^{(u)})^{-2} \cdot (\tilde{\mathbf{A}}_k(\phi_l))'\right)\bigg|_{\phi_l=0} \\
&\geq 2\,\mathrm{Tr}\left(\sum_{i=1}^{c}\mathfrak{p}_i\mathfrak{q}_i^\top \cdot (\tilde{\mathbf{A}}^{(u)})^{-1}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) - \min\lambda(\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}) \cdot \max\lambda((\tilde{\mathbf{A}}^{(u)})^{-2}) \cdot \mathrm{Tr}\left((\tilde{\mathbf{A}}_k(\phi_l))'\right)\bigg|_{\phi_l=0} \\
&= 2\,\mathrm{Tr}\left(\sum_{i=1}^{c}\mathfrak{p}_i\mathfrak{q}_i^\top \cdot (\tilde{\mathbf{A}}^{(u)})^{-1}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) - \min\lambda(\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}) \cdot (\min\lambda(\tilde{\mathbf{A}}^{(u)}))^2 \cdot \mathrm{Tr}\left((\tilde{\mathbf{A}}_k(\phi_l))'\right)\bigg|_{\phi_l=0} \\
&\geq 2\,\mathrm{Tr}\left(\sum_{i=1}^{c}\mathfrak{p}_i\mathfrak{q}_i^\top \cdot (\tilde{\mathbf{A}}^{(u)})^{-1}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) - \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \cdot (\min\lambda(\tilde{\mathbf{A}}^{(u)}))^2 \cdot \mathrm{Tr}\left((\tilde{\mathbf{A}}_k(\phi_l))'\right)\bigg|_{\phi_l=0} \\
&\geq 2\,\mathrm{Tr}\left(\sum_{i=1}^{c}\mathfrak{p}_i\mathfrak{q}_i^\top \cdot (\tilde{\mathbf{A}}^{(u)})^{-1}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) - \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \cdot \|\tilde{\mathbf{A}}^{(u)}\|_F^2 \cdot \mathrm{Tr}\left((\tilde{\mathbf{A}}_k(\phi_l))'\right)\bigg|_{\phi_l=0} \\
&= 2\,\mathrm{Tr}\left(\sum_{i=1}^{c}\mathfrak{p}_i\mathfrak{q}_i^\top \tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) - \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2 \cdot \|\tilde{\mathbf{A}}^{(u)}\|_F^2 \cdot \mathrm{Tr}\left((\tilde{\mathbf{A}}_k(\phi_l))'\right)\bigg|_{\phi_l=0}
\end{aligned}
\tag{77}
$$

The first two inequalities are again the inequality for the trace of matrix product of two square matrices (Fang et al., 1994). The third inequality holds because:

$$
\min\lambda(\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}) < \mathrm{Tr}(\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}) = \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2.
\tag{78}
$$

The last inequality holds because:

$$
(\min\lambda(\tilde{\mathbf{A}}^{(u)}))^2 = \min\lambda(\tilde{\mathbf{A}}^{(u)2}) = \min\lambda(\tilde{\mathbf{A}}^{(u)} \cdot \tilde{\mathbf{A}}^{(u)\top}) \leq \|\tilde{\mathbf{A}}^{(u)}\|_F^2
\tag{79}
$$

Finally, the last equation holds true because of Assumption 2 where the perturbation vector $\mathfrak{q}_i$ lies in the linear span of the matrix $\mathbf{V}_k^{(u)}[\mathbf{\Sigma}_k^{(u)}]^{-\frac{1}{2}}$:

$$
\mathfrak{q}_i^\top \cdot (\tilde{\mathbf{A}}^{(u)})^{-1} = \mathfrak{q}_i^\top \cdot \mathbf{V}_k^{(u)}\mathbf{\Sigma}_k^{(u)-1}\mathbf{V}_k^{(u)\top} = \mathfrak{q}_i^\top.
\tag{80}
$$

For the key term $\mathrm{Tr}\left(\tilde{\mathbf{A}}_k(\phi_l)'\right)\bigg|_{\phi_l=0}$, it is easy to check that $\frac{\lambda_j^{(u)}}{\lambda_j^{(u)}-\lambda_i^{(u)}} \leq \frac{\tau}{\tau-1}, j \leq k < k+1 \leq i$ because τ is a constant that measures the k-th spectral gap of matrix $\tilde{\mathbf{A}}^{(u)}$, i.e., $\lambda_k^{(u)} \geq \tau\lambda_{k+1}^{(u)}$, we can similarly get its upper bound to Equation 73

as follows:

$$
\begin{aligned}
\operatorname{Tr}\left(\tilde{\mathbf{A}}_k(\phi_l)'\right)\Big|_{\phi_l=0} &= \operatorname{Tr}\left(\left(\mathbf{V}_k(\phi_l)\boldsymbol{\Sigma}_k(\phi_l)\mathbf{V}_k(\phi_l)^\top\right)'\right)\Big|_{\phi_l=0} \\
&\leq \Bigg[\operatorname{Tr}(\mathbf{V}_k^{(u)}\cdot\mathbf{V}_k^{(u)\top}\sum_{i=1}^{c}\mathfrak{q}_i\mathfrak{q}_i{}^\top\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}) - 2\operatorname{Tr}(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\mathbf{V}_k^{(u)}\cdot\mathbf{V}_k^{(u)\top}) \\
&\quad + 2\frac{\tau}{\tau-1}\operatorname{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^{c}\mathfrak{q}_i\mathfrak{q}_i{}^\top\left(\mathbf{I}_N-\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\right)\right) - 2\frac{\tau}{\tau-1}\operatorname{Tr}\left(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\left(\mathbf{I}_N-\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\right)\right)\Bigg] \\
&= 2\frac{\tau}{\tau-1}\operatorname{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^{c}\mathfrak{q}_i\mathfrak{q}_i{}^\top\right) - \frac{\tau+1}{\tau-1}\operatorname{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^{c}\mathfrak{q}_i\mathfrak{q}_i{}^\top\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\right) \\
&\quad + \frac{2}{\tau-1}\operatorname{Tr}\left(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\right) - 2\frac{\tau}{\tau-1}\operatorname{Tr}\left(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\right) \\
&= \operatorname{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^{c}\mathfrak{q}_i\mathfrak{q}_i{}^\top\right) + \frac{2}{\tau-1}\operatorname{Tr}\left(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\right) - 2\frac{\tau}{\tau-1}\operatorname{Tr}\left(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\right) \\
&\leq \operatorname{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^{c}\mathfrak{q}_i\mathfrak{q}_i{}^\top\right) + \frac{2}{\tau-1}\cdot\|\mathbf{V}_k^{(u)}\|_F^2\operatorname{Tr}\left(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\right) - 2\frac{\tau}{\tau-1}\operatorname{Tr}\left(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\right) \\
&\leq \operatorname{Tr}\left(\mathbf{V}_k^{(u)}\mathbf{V}_k^{(u)\top}\sum_{i=1}^{c}\mathfrak{q}_i\mathfrak{q}_i{}^\top\right) + \frac{2(k-\tau)}{\tau-1}\operatorname{Tr}\left(\tilde{\mathbf{A}}_k^{(u)}\mathbf{D}_l\right) \\
&= \sum_{i=1}^{c}\|\mathfrak{q}_i\|_F^2 - \frac{2\mathbf{F}_i^{(u)\top}\mathbf{F}_i^{(u)}(\tau-k)}{\tau-1}\sum_{i=1}^{c}\|\mathfrak{q}_i\|_1 = \sum_{i=1}^{c}\|\mathfrak{q}_i\|_F^2 - \frac{2r^2(\tau-k)}{\tau-1}\sum_{i=1}^{c}\|\mathfrak{q}_i\|_1
\end{aligned}
\tag{81}
$$

Therefore, it is natural to obtain that:

$$
\begin{aligned}
&\operatorname{Tr}\left(\left(\tilde{\mathbf{A}}_{\mathrm{OI}}(\phi_l)\mathbf{V}_k(\phi_l)\boldsymbol{\Sigma}_k(\phi_l)^{-1}\mathbf{V}_k(\phi_l)^\top\tilde{\mathbf{A}}_{\mathrm{OI}}(\phi_l)^\top\right)'\right)\Bigg|_{\phi_l=0} \geq \\
&2\sum_{i=1}^{c}\operatorname{Tr}\left(\mathfrak{p}_i\mathfrak{q}_i^\top\cdot\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)\top}\right) - \|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F^2\|\tilde{\mathbf{A}}^{(u)}\|_F^2\cdot\left(\sum_{i=1}^{c}\|\mathfrak{q}_i\|_F^2 - \frac{2r^2(\tau-k)}{\tau-1}\sum_{i=1}^{c}\|\mathfrak{q}_i\|_1\right).
\end{aligned}
\tag{82}
$$

$\square$

E. Empirical Verification of the Main Theorems

Verification of Theorems. We provide more verification results on CIFAR10. Firstly, we verify how the value of $\mathcal{G}$ will change given a larger Frobenius norm of $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}$.

OOD dataset	SVHN	C100
	FAR OOD	NEAR OOD
$\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F \uparrow$	2583	2876
$\mathcal{G} \uparrow$	0.00	**0.14**

Table 5: Verification with different $\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F$ (CIFAR10 as ID).

Next we verify the relationship of $\|\tilde{\mathbf{A}}^{(u)}\|_F$ and the error difference (CIFAR100 as OOD) in Table 6.

Epochs	40	80	120	160	200	240
$\|\tilde{\mathbf{A}}^{(u)}\|_F \downarrow$	19873	19762	17539	16640	15982	15361
$\mathcal{G} \uparrow$	0.03	0.04	0.07	0.11	0.14	0.15

Table 6: Verification with different $\|\tilde{\mathbf{A}}^{(u)}\|_F$ (CIFAR10 dataset as ID).

These two tables show a similar result as when the ID dataset is CIFAR100, where the error difference on CIFAR100 (near OOD, with larger $\|\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}\|_F$) is consistently larger than that on SVHN (far OOD). Moreover, the difference in linear probing error $\mathcal{G}$ tends to increase with decreasing $\|\tilde{\mathbf{A}}^{(u)}\|_F$, which aligns with Theorem 2.

Verification of the assumptions. Here are the empirical verifications of the assumptions on real-world datasets, i.e., CIFAR10 and CIFAR100 datasets and simulated datasets as shown in Figure 2.

For CIFAR10 and CIFAR100, we have checked the largest eigengap of matrix $\tilde{\mathbf{A}}^{(u)}$ and observed that $\tau = 7740.92$ for CIFAR10 and $\tau = 8834.78$ for CIFAR100, which are much larger than the feature dimension 512. Therefore, we can always find a proper k such that $\tau > k$ is satisfied because $k \leq 512$.

For the simulated datasets as shown in Figure 2, the largest eigengap of matrix $\tilde{\mathbf{A}}^{(u)}$ is 872.00. Therefore the condition $\tau > k$ always holds because $k = 2$ in this case. We have provided more details about the matrix itself in Section F.

F. Details of the Illustrative Example

For Figure 2 in the main paper, we generate the augmentation graph $\mathcal{T}$ as follows:

$$\mathcal{T} = \begin{pmatrix} \mathbf{B}_1 & \mathbf{B}_2 & \mathbf{B}_2 \\ \mathbf{B}_2 & \mathbf{B}_3 & \mathbf{B}_2 \\ \mathbf{B}_2 & \mathbf{B}_2 & \mathbf{B}_4 \end{pmatrix}, \tag{83}$$

where the block-wise matrices $\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3, \mathbf{B}_4 \in \mathbb{R}^{40 \times 40}$ are square matrices. Specifically, $\mathbf{B}_1$ has the following definition:

$$\begin{aligned} \mathbf{B}_1 &= \mathbf{1}_{40 \times 40} \cdot B_1 + \frac{1}{2}\varepsilon_1 + \frac{1}{2}\varepsilon_1^\top, \\ \mathbf{B}_2 &= \mathbf{1}_{40 \times 40} \cdot B_2 + \frac{1}{2}\varepsilon_2 + \frac{1}{2}\varepsilon_2^\top, \\ \mathbf{B}_3 &= \mathbf{1}_{40 \times 40} \cdot B_3 + \frac{1}{2}\varepsilon_3 + \frac{1}{2}\varepsilon_3^\top, \\ \mathbf{B}_4 &= \mathbf{1}_{40 \times 40} \cdot B_4 + \frac{1}{2}\varepsilon_4 + \frac{1}{2}\varepsilon_4^\top, \end{aligned} \tag{84}$$

here $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4 \in \mathbb{R}^{40 \times 40}$ are matrices where each element of them is sampled from a truncated normal distribution (lower bound is -0.1, upper bound is 0.1, mean is 0 and variance is 1). $\mathbf{1}_{40 \times 40}$ is a matrix where each element in it is 1. Essentially, the matrix $\mathbf{B}_1$ measures the connectivity pattern between the data that belongs to the first ID class. We set $B_1 = 0.8$.

Similarly, $\mathbf{B}_3$ and $\mathbf{B}_4$ measure the connectivity pattern between the data that belongs to the second/third ID class, and we set $B_3 = 0.75$, $B_4 = 0.7$ respectively for $\mathbf{B}_3$ and $\mathbf{B}_4$. Moreover, the matrix $\mathbf{B}_2$ measures the connectivity pattern between the data that belongs to different ID classes and we set $B_2 = 0.1$ in this case. For the ID adjacency matrix $\mathbf{A}^{(u)}$ and $\mathbf{A}^{(l)}$, we can follow the definition in Definitions 3 and 4 of the main paper for calculation.

In addition, we generate the OOD-ID adjacency matrix $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(u)}$ by sampling from a truncated normal distribution where the lower bound is 0, the upper bound is 0.5, the mean is 0.5 and variance is 0.05 in the near OOD scenario. In the far OOD scenario, we set the lower bound to be 0, the upper bound to be 0.2, the mean to be 0.2, and the variance to be 0.05 for the truncated normal distribution. In the labeled case, the OOD-ID adjacency matrix $\tilde{\mathbf{A}}_{\mathrm{OI}}^{(l)}$ can be calculated according to Equation 23. Here $\mathfrak{q}$ can be calculated based on the augmentation graph $\mathcal{T}$ and each element of $\mathfrak{p}$ (that represents the connectivity probability between the OOD and ID data) is set to 0.1 because we let the connectivity probability be 0.1 when the data belongs to different classes in the augmentation graph $\mathcal{T}$. Finally, we set $\phi_u = 1$ and $\phi_l = 0.5$ for the calculation of all the adjacency matrices and the data representations.

G. Additional Results on the Illustrative Example

In this section, we provide additional visualization results on changing the Frobenius norm of the ID adjacency matrix $\tilde{\mathbf{A}}^{(u)}$ and the semantic connection $\mathfrak{q}$ in Figure 3. For all the examples, we visualize under the near OOD scenario (meaning that we use the same $\tilde{\mathbf{A}}_{\mathrm{OI}}$ as in Section F). Please check the caption for a detailed explanation.
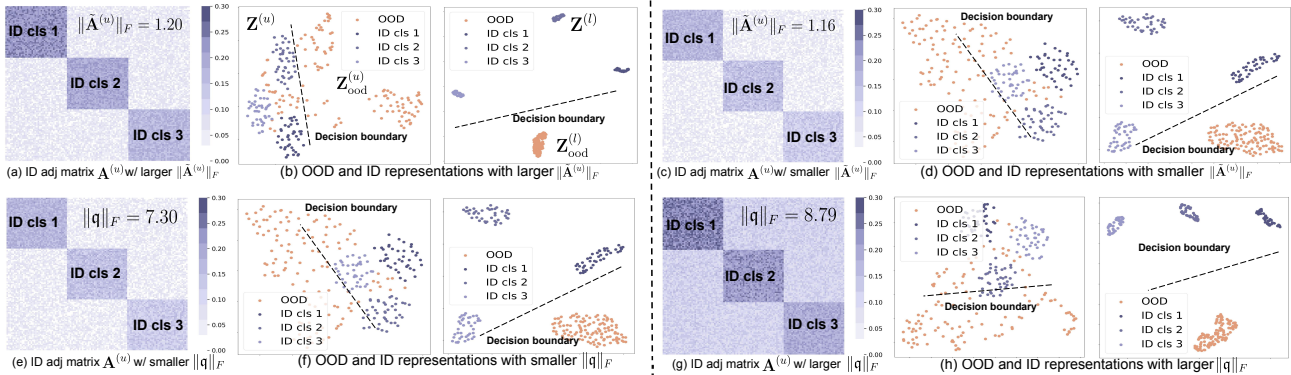

Figure 3: Additional example showcasing the contrast between adjacency matrices and representations w/ (l) and w/o (u) ID labels. (a) The ID adjacency matrix in the unlabeled case $\mathbf{A}^{(u)}$ with a larger $\|\tilde{\mathbf{A}}^{(u)}\|_F$ ($B_1 = 0.8, B_2 = 0.1, B_3 = 0.75, B_4 = 0.7$). (b) The contrast of the learned representations in both labeled and unlabeled cases when $\|\tilde{\mathbf{A}}^{(u)}\|_F = 1.20$. (c) The ID adjacency matrix in the unlabeled case $\mathbf{A}^{(u)}$ with a smaller $\|\tilde{\mathbf{A}}^{(u)}\|_F$ ($B_1 = 0.7, B_2 = 0.1, B_3 = 0.65, B_4 = 0.6$). (d) The contrast of the learned representations in both labeled and unlabeled cases when $\|\tilde{\mathbf{A}}^{(u)}\|_F = 1.16$. Compared with (b) where the difference in the linear probing loss $\mathcal{G}$ is 0.09, the linear probing loss reduces from 0.16 to 0.02. (e) The ID adjacency matrix in the unlabeled case $\mathbf{A}^{(u)}$ with a smaller $\|\mathfrak{q}\|_F$ ($B_1 = 0.7, B_2 = 0.1, B_3 = 0.65, B_4 = 0.6$). (f) The contrast of the learned representations in both labeled and unlabeled cases when $\|\mathfrak{q}\|_F = 7.30$. (g) The ID adjacency matrix in the unlabeled case $\mathbf{A}^{(u)}$ with a larger $\|\mathfrak{q}\|_F$ ($B_1 = 0.8, B_2 = 0.2, B_3 = 0.75, B_4 = 0.7$). (h) The contrast of the learned representations in both labeled and unlabeled cases when $\|\mathfrak{q}\|_F = 8.79$. Compared with (f) where the difference in the linear probing loss $\mathcal{G}$ is 0.14, the linear probing loss reduces from 0.16 to 0.00. The visualization aligns with our theoretical reasoning as shown in Section 4.3.

H. Additional Experimental Details

We provide more training details for both contrastive learning and linear probing. For contrastive training, we use the same data augmentation strategies as SimSiam (Chen & He, 2021). For CIFAR10, we set $\phi_u = 0.5, \phi_l = 0.25$ with training epoch 200, and we evaluate using features extracted from the layer preceding the projection. For CIFAR100, we set $\phi_u = 3, \phi_l = 0.0225$ with 200 training epochs and assess based on the projection layer's features. We use SGD with momentum 0.9 as an optimizer with cosine annealing (lr=0.03), weight decay 5e-4, and batch size 512.

For linear probing, we train a linear layer on the extracted features from the pretrained model by contrastive learning. We use SGD for 50 epochs with momentum 0.9 as an optimizer with a decayed learning rate in epoch 30 by 0.2 (The initial learning rate is 5), and batch size 512.

I. Additional Experimental Results

Results on CIFAR10. We present the experimental results on CIFAR10 in Table 7, where the effect of the ID labels is similar to the results on CIFAR100.

OOD category	OOD dataset	ID labels	FPR95	AUROC	LP error	FPR95	AUROC	LP error
			$\mathbb{P}_{ood}^{test} = \mathbb{P}_{ood}^{LP}$			$\mathbb{P}_{ood}^{test} \neq \mathbb{P}_{ood}^{LP}$		
FAR OOD	SVHN	-	$0.05^{\pm0.01}$	$\mathbf{99.98}^{\pm0.00}$	$0.01^{\pm0.00}$	$36.60^{\pm2.84}$	$91.71^{\pm1.90}$	$0.33^{\pm0.11}$
		+	$\mathbf{0.03}^{\pm0.00}$	$99.98^{\pm0.01}$	$\mathbf{0.00}^{\pm0.00}$	$\mathbf{8.62}^{\pm1.12}$	$\mathbf{98.43}^{\pm0.47}$	$\mathbf{0.25}^{\pm0.19}$
	TEXTURES	-	$\mathbf{0.21}^{\pm0.03}$	$\mathbf{99.96}^{\pm0.01}$	$0.01^{\pm0.00}$	$\mathbf{10.16}^{\pm0.72}$	$\mathbf{98.13}^{\pm2.03}$	$\mathbf{0.52}^{\pm0.10}$
		+	$0.57^{\pm0.06}$	$99.80^{\pm0.10}$	$0.01^{\pm0.00}$	$13.87^{\pm0.82}$	$97.60^{\pm1.39}$	$0.58^{\pm0.18}$
	PLACES365	-	$\mathbf{0.36}^{\pm0.01}$	$\mathbf{99.91}^{\pm0.00}$	$0.02^{\pm0.00}$	$20.19^{\pm1.23}$	$96.18^{\pm0.28}$	$0.43^{\pm0.16}$
		+	$0.68^{\pm0.02}$	$99.68^{\pm0.06}$	$0.02^{\pm0.01}$	$\mathbf{15.83}^{\pm1.83}$	$\mathbf{97.01}^{\pm0.96}$	$\mathbf{0.26}^{\pm0.06}$
	LSUN-RESIZE	-	$0.68^{\pm0.01}$	$99.78^{\pm0.18}$	$0.03^{\pm0.00}$	$18.21^{\pm1.71}$	$96.87^{\pm3.28}$	$0.42^{\pm0.21}$
		+	$\mathbf{0.56}^{\pm0.23}$	$99.78^{\pm0.10}$	$\mathbf{0.01}^{\pm0.01}$	$\mathbf{10.83}^{\pm1.18}$	$\mathbf{97.94}^{\pm0.23}$	$\mathbf{0.39}^{\pm0.03}$
	LSUN-C	-	$0.68^{\pm0.12}$	$99.80^{\pm0.05}$	$0.04^{\pm0.02}$	$11.10^{\pm1.78}$	$97.93^{\pm1.89}$	$0.42^{\pm0.07}$
		+	$\mathbf{0.32}^{\pm0.16}$	$99.65^{\pm0.05}$	$\mathbf{0.02}^{\pm0.02}$	$\mathbf{7.95}^{\pm1.92}$	$\mathbf{98.44}^{\pm0.21}$	$\mathbf{0.40}^{\pm0.19}$
NEAR OOD	CIFAR100	-	$49.88^{\pm1.81}$	$89.29^{\pm0.99}$	$0.24^{\pm0.07}$	$54.76^{\pm2.21}$	$86.65^{\pm3.04}$	$0.53^{\pm0.08}$
		+	$\mathbf{41.56}^{\pm1.49}$	$\mathbf{92.83}^{\pm0.73}$	$\mathbf{0.10}^{\pm0.02}$	$\mathbf{40.04}^{\pm2.92}$	$\mathbf{92.42}^{\pm0.18}$	$\mathbf{0.41}^{\pm0.19}$

Table 7: OOD detection results w/ and w/o ID labels (CIFAR10 as ID). Mean and std are estimated on three different runs. Better results are highlighted in bold. "+,-" denotes the labeled and unlabeled case. "LP error" denotes the error of linear probing.

Results on using post-hoc OOD detection scores. Instead of using linear probing to evaluate the OOD detection performance, we investigate another approach, which directly calculates the k-NN score (Sun et al., 2022) on top of the extracted representations for both the ID and OOD data and then computes the OOD detection metrics for comparison. The result is shown in Table 8, where the OOD detection performance is usually better in the labeled case.

OOD category	OOD dataset	ID labels	FPR95	AUROC	FPR95	AUROC
			CIFAR10 as ID		CIFAR100 as ID	
FAR OOD	SVHN	-	$37.35^{\pm3.12}$	$88.08^{\pm2.91}$	$\mathbf{27.27}^{\pm0.88}$	$\mathbf{93.85}^{\pm2.16}$
		+	$\mathbf{14.97}^{\pm1.81}$	$\mathbf{97.82}^{\pm0.96}$	$55.39^{\pm4.78}$	$86.31^{\pm1.12}$
	TEXTURES	-	$63.91^{\pm2.16}$	$78.29^{\pm3.06}$	$77.87^{\pm1.97}$	$61.68^{\pm1.68}$
		+	$\mathbf{44.91}^{\pm2.90}$	$\mathbf{92.89}^{\pm0.95}$	$\mathbf{67.19}^{\pm2.08}$	$\mathbf{81.11}^{\pm2.89}$
	PLACES365	-	$65.90^{\pm4.02}$	$86.91^{\pm3.38}$	$76.52^{\pm4.29}$	$70.68^{\pm2.90}$
		+	$\mathbf{61.02}^{\pm2.61}$	$\mathbf{92.65}^{\pm1.91}$	$\mathbf{64.16}^{\pm0.72}$	$\mathbf{85.28}^{\pm1.29}$
	LSUN-RESIZE	-	$68.91^{\pm2.17}$	$85.89^{\pm0.90}$	$90.00^{\pm1.38}$	$57.91^{\pm3.96}$
		+	$\mathbf{56.81}^{\pm0.61}$	$\mathbf{92.81}^{\pm1.74}$	$\mathbf{67.40}^{\pm3.19}$	$\mathbf{85.89}^{\pm3.77}$
	LSUN-C	-	$\mathbf{26.08}^{\pm1.48}$	$93.21^{\pm0.27}$	$62.57^{\pm0.97}$	$75.50^{\pm1.98}$
		+	$36.32^{\pm2.20}$	$\mathbf{94.97}^{\pm1.88}$	$\mathbf{57.10}^{\pm2.93}$	$\mathbf{83.23}^{\pm2.26}$
NEAR OOD	CIFAR100/CIFAR10	-	$71.76^{\pm2.49}$	$81.97^{\pm0.71}$	$92.10^{\pm2.90}$	$58.57^{\pm4.86}$
		+	$\mathbf{53.22}^{\pm2.10}$	$\mathbf{90.94}^{\pm1.17}$	$\mathbf{84.12}^{\pm2.26}$	$\mathbf{74.31}^{\pm0.63}$

Table 8: OOD detection results measured by post-hoc k-NN score w/ and w/o ID labels (CIFAR10 and CIFAR100 as ID). Mean and std are estimated on three different runs. Better results are highlighted in bold. "+,-" denotes the labeled and unlabeled case and k is set to 25 for all the experiments.

J. Tightness of the Bound

We provide the evidence to verify the tightness of our bound. Specifically, we present numerical results on the illustrative example (Figure 2) to show the proximity between the value of the error difference $\mathcal{G}$ and the bound in our Theorem in Table 9. Specifically, we set a different value of $\|\tilde{\mathbf{A}}_{OI}^{(u)}\|_F$ and observed our lower bound is sufficiently close to the error difference $\mathcal{G}$ (The details of the dataset used are the same as those described in Appendix Section F except for the Frobenius norm of the OOD-ID adjacency matrix).

$\|\tilde{\mathbf{A}}_{OI}^{(u)}\|_F$	$\mathcal{G}$	Our bound
60	0.09	0.07
72	0.16	0.12
84	0.21	0.16
96	0.39	0.37
108	0.40	0.34
120	0.61	0.56

Table 9: Numerical results on the bound tightness.