A Proofs

A.1 Proof of Theorem 4.5

Proof. We first rewrite $\mathcal{L}_{\text{tri}}$ as a matrix decomposition objective

$$\begin{aligned}\mathcal{L}_{\text{tri}}(f) &= -2\mathbb{E}_{x,x^+} f(x)^\top S f(x^+) + \mathbb{E}_x \mathbb{E}_{x^-} \left(f(x)^\top S f(x^-)\right)^2 \\ &= \sum_{x,x'} \left(\frac{A_{xx'}^2}{D_{xx} D_{x'x'}} + D_{xx} D_{x'x'} \left(f(x)^\top S f(x')\right)^2 - 2A_{xx'} f(x)^\top S f(x') \right) + const \quad (16) \\ &= \|\bar{A} - FSF^\top\|^2.\end{aligned}$$

According to the Eckart-Young Theorem [Eckart and Young, 1936], the optimal solutions $F^\star, S^\star$ satisfy

$$F^\star S^\star (F^\star)^\top = U^k \Sigma (V^k)^\top,$$

where $\Sigma \in \mathbb{R}^{k\times k}$ is a diagonal matrix with the k-largest eigenvalues of $\bar{A}$ and $U \in \mathbb{R}^{N\times k}$ contains the corresponding eigenvectors of the k-largest eigenvalues. When the regularizer $\mathcal{L}_{\text{Dec}}$ is minimized, $F^\star$ satisfy $(F^\star)^\top F^\star = I$. In the next step, we prove the uniqueness of the optimal solution.

We denote $H = F^\star \Sigma (F^\star)^\top$. As $(F^\star)^\top F^\star = I$, we obtain $HH^\top = F^\star S^\star (S^\star)^\top (F^\star)^\top$. If ζ, σ are a pair of eigenvector and eigenvalue of $HH^\top$, we have

$$\begin{aligned} HH^\top \zeta = F^\star S^\star (S^\star)^\top (F^\star)^\top \zeta &= \sigma\zeta, \\ S^\star (S^\star)^\top (F^\star)^\top \zeta &= \sigma (F^\star)^\top \zeta, \\ S^\star (S^\star)^\top \left((F^\star)^\top \zeta\right) &= \sigma \left((F^\star)^\top \zeta\right). \end{aligned} \quad (17)$$

So the eigenvalues of $HH^\top$ are the eigenvalues of $S^\star (S^\star)^\top$. As the positive eigenvalues of $HH^\top$ are uniquely determined and $S^\star$ has a descending order, $S^\star$ is also determined and $S^\star = \sum$.

We note that $HH^\top = F^\star S^\star (S^\star)^\top (F^\star)^\top$, i.e., $HH^\top F^\star = F^\star S^\star (S^\star)^\top$, which means that the k columns of $F^\star$ are the eigenvectors of $HH^\top$ and the corresponding eigenvalues are $\sigma_1 \cdots \sigma_k$. As $HH^\top$ only has k different non-negative eigenvalues $\sigma_1, \cdots, \sigma_k$, the eigenspace of each eigenvalue is one-dimensional. When we consider the real number space, any two eigenvectors ζ_i, ζ_i' of the same eigenvalue σ_i satisfy $\zeta_i = c\zeta_i'$. As $(F^\star)^\top F^\star = I$, we obtain $c = \pm 1$. As $f(x) = \frac{1}{\sqrt{D_{xx}}} F_x$, we obtain

$$f_j^\star(x) = \pm \frac{1}{\sqrt{D_{xx}}} \left(U_x^k\right)_j, S^* = \text{diag}(\sigma_1, \ldots, \sigma_k), \quad (18)$$

□

A.2 Proof of Theorem 5.1

We first introduce a lemma which theoretically guarantees the generalization performance of spectral contrastive learning.

Lemma A.1 ([HaoChen et al., 2021])**.** *For the optimal solutions to spectral contrastive learning (SCL), we have*

$$\mathcal{E}(f_{SCL}^\star) \leq \mathcal{O}(\frac{\alpha}{1-\sigma_{k+1}}),$$

where we denote α as the probability that the natural samples and augmented views have different labels, i.e., $\alpha = \mathbb{E}_{\bar{x}\sim\mathcal{P}_u} \mathbb{E}_{x\sim\mathcal{A}(\cdot|\bar{x})} \mathbb{1}[y(\bar{x}) \neq y(x)]$ and σ_{k+1} as the $(k+1)$-th largest eigenvalue of the normalized adjacent matrix $\bar{A}$.

Then we construct the generalization guarantee of tri-contrastive learning.

Proof. Following the proof of Theorem 4.5, we know that the optimal solutions learned by triCL are

$$\begin{aligned} F^\star &= U^k, \\ S^\star &= diag(\sigma_1, \cdots \sigma_k). \end{aligned}$$

So we know that the optimal encoder of triCL satisfies, $\forall x \in \mathcal{D}$

$$f^*(x) = \frac{1}{\sqrt{D_{xx}}} \left(U_x^k\right)^\top.$$

Compared with the optimal solutions of spectral contrastive learning (Eq. 2), we know

$$(\mathrm{diag}(\sigma_1, \cdots \sigma_k)R)^\top f^*_{triCL}(x) = f^\star_{SCL}(x), \tag{19}$$

where $f^\star_{triCL}, f^\star_{SCL}$ denote the optimal solutions of tri-contrastive learning and spectral contrastive learning. As $\mathrm{diag}(\sigma_1, \cdots \sigma_k)R$ is an invertible matrix, we then prove that the invertible matrix can be absorbed in the linear probing. We denote $\mathrm{diag}(\sigma_1, \cdots \sigma_k)R$ as Q and we denote the linear classifier as B, i.e., $g(f(x)) = f(x)^\top B$. For a linear classfier B, let $\tilde{B} = BQ^{-1}$. We then obtain $f^\star_{triCL}(x)^\top \tilde{B} = f^\star_{SCL}(x)^\top B$.

So

$$\mathcal{E}(f^\star_{triCL}) = \mathcal{E}(f^\star_{SCL}).$$

With lemma A.1, we have

$$\mathcal{E}(f^\star_{triCL}) \leq \mathcal{O}(\frac{\alpha}{1-\sigma_{k+1}}).$$

□

A.3 Proof of Theorem 5.2

Proof. Based on the proof of Theorem 4.5, we know that the t-th dimension of the optimal solutions satisfies

$$F^\star = U_t^k,$$
$$S_t^\star = diag(\sigma_1, \cdots \sigma_k)_t.$$

With the analysis in Eckart-Young theorem [Eckart and Young, 1936], we have

$$\begin{aligned}\|\bar{A} - F_t^\star S_t^\star (F_t^\star)^\top\|_F^2 &= \|\bar{A} - U_t^k \,\mathrm{diag}(\sigma_1, \cdots \sigma_k)_t (U_t^k)^\top\|_F^2 \\ &= \sum_{i=1}^{t-1} \sigma_i^2 + \sum_{i=t+1}^{k} \sigma_i^2.\end{aligned}$$

As σ_i is the i-th largest eigenvalues of $\bar{A}$, so

$$\|\bar{A} - F_1^\star S_1^\star (F_1^\star)^\top\|_F^2 \leq \cdots \leq \|\bar{A} - F_k^\star S_k^\star (F_k^\star)^\top\|_F^2.$$

Following Eq 16, we obtain

$$\mathcal{L}_{triCL}(f_t, S_t) = \|\bar{A} - F_t^\star S_t^\star (F_t^\star)^\top\|_F^2 + const,$$

we obtain

$$\mathcal{L}_{\mathrm{triCL}}(f_1^\star, S_1^\star) \leq \cdots \leq \mathcal{L}_{\mathrm{triCL}}(f_k^\star, S_k^\star).$$

□

A.4 Feature Identifiability of Asymmetric Tri-contrastive Learning

We first extend the augmentation graph to an asymmetric form. The asymmetric augmentation graph is defined over the set of all samples with its adjacent matrix denoted by P_O. In the augmentation graph, each node corresponds to a sample, and the weight of the edge connecting two nodes x_A and x_B is equal to the probability that they are selected as a positive pair, i.e.,$(P_O)_{x_a,x_b} = \mathcal{P}_O(x_a, x_b)$. And we denote $\bar{P}_O$ as the normalized adjacent matrix of the augmentation graph, i.e., $(\bar{P}_O)_{x_a,x_b} = \frac{\mathcal{P}_O(x_a,x_b)^2}{\mathcal{P}_A(x_a)\mathcal{P}_B(x_b)}$.

Similar to the symmetric form, we then rewrite $\mathcal{L}_{\text{tri}}$ as a matrix decomposition objective

$$\begin{aligned}\mathcal{L}_{\text{tri}}(f_A, f_B, S) &= -2\mathbb{E}_{x_a,x_b} f_A(x_a)^\top S f_B(x_b) + \mathbb{E}_{x_a^-,x_b^-}\left(f_A(x_a^-)^\top S f_B(x_b^-)\right)^2 \\ &= \sum_{x_a,x_b}\Big(\frac{\mathcal{P}_O(x_a,x_b)^2}{\mathcal{P}_A(x_a)\mathcal{P}_B(x_b)} + \mathcal{P}_A(x_a)\mathcal{P}_B(x_b)\left(f_A(x_a)^\top S f_B(x_L)\right)^2 \\ &- 2\mathcal{P}_O(x_a,x_b) f_A(x_a)^\top S f_B(x_L)\Big) + const \\ &= \|\bar{P}_O - F_A S F_B^\top\|^2.\end{aligned}$$

According to the Eckart-Young Theorem [Eckart and Young, 1936], the optimal solutions $F_A^\star, S^\star, F_B^\star$ satisfy

$$F_A^\star S^\star (F_B^\star)^\top = U^k \Sigma (V^k)^\top,$$

where $\Sigma \in \mathbb{R}^{k\times k}$ is a diagonal matrix with the k-largest eigenvalues of $\bar{P}_O$ and $U \in \mathbb{R}^{N_A\times k}$ contains the corresponding eigenvectors of the k-largest eigenvalues. When the regularizer $\mathcal{L}_{\text{Dec}}$ is minimized, $F_A^\star$ and $F_B^\star$ satisfy $(F_A^\star)^\top F_A^\star = I$, $(F_B^\star)^\top F_B^\star = I$. In the next step, we prove the uniqueness of the optimal solution.

We denote $H = F_A^\star \Sigma F_B^\star$, and we obtain $HH^\top = F_A^\star S^\star (S^\star)^\top (F_A^\star)^\top$. If ζ, σ are a pair of eigenvector and eigenvalue of $HH^\top$, we have

$$\begin{aligned} HH^\top\zeta = F_A^\star S^\star (S^\star)^\top (F_A^\star)^\top \zeta &= \sigma\zeta, \\ S^\star (S^\star)^\top (F_A^\star)^\top \zeta &= \sigma (F_A^\star)^\top \zeta, \\ S^\star (S^\star)^\top \left((F_A^\star)^\top \zeta\right) &= \sigma\left((F_A^\star)^\top \zeta\right). \end{aligned} \tag{20}$$

So the eigenvalues of $HH^\top$ are the eigenvalues of $S^\star(S^\star)^\top$. As the positive eigenvalues of $HH^\top$ are uniquely determined and $S^\star$ has an increasing order, $S^\star$ is also determined and $S^\star = \sum$.

We note that $HH^\top = F_A^\star S^\star (S^\star)^\top (F_A^\star)^\top$, i.e., $HH^\top F_A^\star = F_A^\star S^\star (S^\star)^\top$, which means that the k columns of $F_A^\star$ are the eigenvectors of $HH^\top$ and the corresponding eigenvalues are $\sigma_1 \cdots \sigma_k$. As $HH^\top$ only has k different non-negative eigenvalues $\sigma_1, \cdots, \sigma_k$, the eigenspace of each eigenvalue is one-dimensional. When we consider the real number space, any two eigenvectors ζ_i, ζ_i' of the same eigenvalue σ_i satisfy $\zeta_i = c\zeta_i'$. As $(F_A^\star)^\top F_A^\star = I$, we obtain $c = +1$. Then we eliminate the ambiguity of sign following Eq 11 and $F_A^\star$ is unique. Similarly, $F_B^\star$ is also unique. So the optimal solution of $\mathcal{L}_{\text{triCLIP}}$ is unique.

B Experimental Details

B.1 Experiment Details of Section 6.1

We first generate a random matrix A with size 5000×3000, and make sure that it does not contain multiple eigenvectors (which is easy to satisfy). For the matrix factorization problem $\|A - FG^\top\|_F^2$, we apply off-the-shelf algorithms and repeat this process ten times. We then calculate the mean and variance of the l_2 pairwise distance between the obtained solutions of F. For the trifactorization objective $\|A - FSG^\top\|_F^2$, we use SVD to obtain an initial solution, and apply the sign identification procedure to determine the sign of each eigenvector. Similarly, we also repeat this process ten times and calculate the mean and variance of the l_2 pairwise distance between different solutions.

B.2 Experiment Details of Section 6.2

Pretraining Setups. For different evaluation tasks (k-NN, linear evaluation, image retrieval), we use the same pretrained models. We adopt ResNet-18 as the backbone. For CIFAR-10 and CIFAR-100, the projector is a two-layer MLP with hidden dimension 2048 and output dimension 256. And for ImageNet-100, the projector is a two-layer MLP with hidden dimension 4096 and output dimension 256. We pretrain the models with batch size 256 and weight decay 0.0001. For CIFAR-10 and CIFAR-100, we pretrain the models for 200 epochs. While for ImageNet-100, we pretrain the models for 400 epochs. We use the cosine anneal learning rate scheduler and set the initial learning rate to 0.4 on CIFAR-10, CIFAR-100, and 0.3 on ImageNet-100.

As the importance matrix is learned on the projection layer, we conduct the downstream tasks on the features encoded by the complete networks (containing both the backbones and the projectors).

The Distribution of the Importance Matrix. When observing the distribution of feature importance discovered by the importance matrix S, we first apply the softmax activation functions on the diagonal values of S and sort different rows of S by the descending order of corresponding diagonal values in S. We denote the non-negative ordered diagonal values of S as $(s_1, \cdots, s_k)$. When we present the distribution of them in Figure 2(a), we normalize the diagonal values and obtain $(s_1/\sum_{i=1}^{k} s_i, \cdots, s_k/\sum_{i=1}^{k} s_i)$.

The K-NN Accuracy on Selected Dimensions. For k-NN evaluation on 10 selected dimensions, we do not finetune the models. We sort the dimensions of $f(x)$ by the descending order of corresponding diagonal values in the importance matrix. The k-NN is conducted on the standard split of CIFAR-10, CIFAR-100 and ImageNet-100 and the predicted label of samples is decided by the 10 nearest neighbors.

Linear Evaluation on Selected Dimensions. We train the linear classifier on 20 dimensions of the frozen networks for 30 epochs during the linear evaluation. We set batch size to 256 and weight decay to 0.0001. For triCL, we sort the dimensions by descending order of the importance matrix. And for SCL, we randomly choose 20 dimensions.

C More Extensions of Tri-contrastive Learning

In this section, we apply tri-contrastive learning to another representative contrastive learning objective: the non-contrastive loss [Grill et al., 2020, Chen and He, 2021].

Besides contrastive learning, non-contrastive learning is another popular self-supervised framework that throws the negative samples in contrastive learning and learns the meaningful representations only by aligning the positive pairs. Taking the state-of-the-art algorithm BYOL [Grill et al., 2020] as an example, they use an MSE loss:

$$\mathcal{L}_{\mathrm{MSE}}(f, g) = 2 - 2 \cdot \mathbb{E}_{x.x^+} \frac{g(x)^\top f(x^+)}{\|g(x)\|_2 \cdot \|f(x^+)\|_2}, \tag{21}$$

where $g(x)$ and $f(x)$ are two different networks to avoid the feature collapse. Then we consider adapting the tri-term loss to the non-contrastive learning, i.e.,

$$\mathcal{L}_{\mathrm{triMSE}}(f, g) = 2 - 2 \cdot \mathbb{E}_{x.x^+} \frac{g(x)^\top S f(x^+)}{\|g(x)\|_2 \cdot \|f(x^+)\|_2} + \left\|\mathbb{E}_x g(x) g(x)^\top - I\right\|^2. \tag{22}$$

It is noticed that BYOL utilizes the stop-gradient technique on the target network f and it is updated by exponential moving average. So we only calculate the feature decorrelation loss on the online network g.