TranSG: Transfomer-Based Skeleton Graph Prototype Contrastive Learning with Structure-Trajectory Prompted Reconstruction for Person Re-Identification – Appendix I Theoretical Hypotheses and Analyses

Haocong Rao Chunyan Miao*
LILY Research Center, Nanyang Technological University, Singapore
School of Computer Science and Engineering, Nanyang Technological University, Singapore
{haocong001, ascymiao}@ntu.edu.sg

The proposed graph prototype contrastive learning (GPC) can be formulated as Expectation-Maximization (EM) solutions. In this appendix, we first provide a theoretical EM modeling for GPC to prove its validity and convergence, and then systematically present the relations and differences between the proposed approach and existing contrastive learning paradigms.

Preliminaries. For clarity and convenience, we adopt a more general notation here, which is different from that used in the paper. Suppose that a training set $S = \{\boldsymbol{s}_i\}_{i=1}^{N}$ contains N skeleton sequences, where $\boldsymbol{s}_i \in \mathbb{R}^{F\times K}$, $K = J\times 3$, J is the number of body joints with 3D positions, and F is the sequence length ($i.e.$, number of consecutive skeletons). We first represent each sequence with skeleton graphs by: $G(\boldsymbol{s}_i) = \boldsymbol{x}_i = (\mathcal{G}_1, \cdots, \mathcal{G}_F)$, where $G(\cdot)$ is the *predefined* graph construction function, $\boldsymbol{x}_i = (\mathcal{G}_1, \cdots, \mathcal{G}_F)$ is the consecutive graphs representing the i^{th} skeleton sequence $\boldsymbol{s}_i$, and $\mathcal{G}_t$ denotes the t^{th} graph corresponding to the t^{th} skeleton in $\boldsymbol{s}_i$. In our work, instead of using original skeleton sequences, we utilize their skeleton graph representations as inputs to capture richer skeletal body and motion features. The objective of skeleton graph representation learning is to learn a graph embedding/encoder function f_θ (realized via θ-parameterized neural networks) that maps the skeleton graphs $\boldsymbol{x}_i$ to $\boldsymbol{v}_i \in \mathbb{R}^H$, by $\boldsymbol{v}_i = f_\theta(\boldsymbol{x}_i)$, such that $\boldsymbol{v}_i$ can effectively represent latent features of $\boldsymbol{x}_i$ to perform person re-identification.

Formally, the goal is to find the network parameter θ that maximizes the log-likelihood function of the observed graph representations $\{\boldsymbol{x}_i\}_{i=1}^{N}$ of N skeleton sequences as follows:

$$\begin{aligned}\theta^* &= \arg\max_\theta L(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N; \theta)\\ &= \arg\max_\theta \prod_{i=1}^{N} p(\boldsymbol{x}_i; \theta)\\ &\iff \arg\max_\theta \sum_{i=1}^{N} \log p(\boldsymbol{x}_i; \theta),\end{aligned} \tag{1}$$

where $L(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N; \theta)$ denotes the likelihood function of the observed skeleton graph representations *with regard to* θ, and each skeleton graph representation[1] $\boldsymbol{x}_i$ is hypothetically related to a certain skeleton graph prototype $\boldsymbol{c}_j \in \mathbb{R}^H$, with $\boldsymbol{c}_j \in \{\boldsymbol{c}_j\}_{j=1}^{K}$ and K is the number of graph prototypes. Under this assumption, we can re-formulate the objective in Eq. (1) as:

$$\begin{aligned}\theta^* =& \arg\max_\theta \sum_{i=1}^{N} \log p(\boldsymbol{x}_i; \theta)\\ =& \arg\max_\theta \sum_{i=1}^{N} \log \sum_{j=1}^{K} p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta),\end{aligned} \tag{2}$$

Directly optimizing this function is intractable, thus we consider a lower-bound by using a surrogate function as:

$$\begin{aligned}&\sum_{i=1}^{N} \log \sum_{j=1}^{K} p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta)\\ =& \sum_{i=1}^{N} \log \sum_{j=1}^{K} Q(\boldsymbol{c}_j) \frac{p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta)}{Q(\boldsymbol{c}_j)}\\ \geq& \sum_{i=1}^{N} \sum_{j=1}^{K} Q(\boldsymbol{c}_j) \log \frac{p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta)}{Q(\boldsymbol{c}_j)},\end{aligned} \tag{3}$$

*Corresponding author

[1] For simplicity of presentation, we use the "skeleton graph representation" to denote the graph representation of a skeleton sequence.

where $Q(\boldsymbol{c}_j)$ represents some distribution over $\{\boldsymbol{c}_j\}_{j=1}^K$ and $\sum_{j=1}^K Q(\boldsymbol{c}_j) = 1$. We apply Jensen's inequality to derive the last step of Eq. (3), where the equality can be achieved under the condition that $\frac{p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta)}{Q(\boldsymbol{c}_j)}$ is a constant. To realize this equality, we have:

$$Q(\boldsymbol{c}_j) = \frac{p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta)}{\sum_{m=1}^K p(\boldsymbol{x}_i, \boldsymbol{c}_m; \theta)} = \frac{p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta)}{p(\boldsymbol{x}_i; \theta)} = p(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta), \quad (4)$$

where $Q(\boldsymbol{c}_j)$ is a posterior probability related to $\boldsymbol{c}_j, \boldsymbol{x}_i$, and θ. Different from [1, 2] that employ clustering to estimate prototypes and representation distributions under the fixed θ, we exploit the graph feature centroid of each *ground-truth* identity as a different graph prototype. In particular, when given the θ-parameterized encoder to encode all skeleton graphs ($\boldsymbol{x}_i$) *at the Expectation step*, their corresponding graph prototypes ($\boldsymbol{c}_j$) are assumed to follow the distribution of ground-truth classes in the dataset. The graph prototype distribution is hence constant and can be computed by $Q(\boldsymbol{c}_j) = p(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta)$. We can re-write Eq. (3) as:

$$\sum_{i=1}^N \sum_{j=1}^K \left(Q(\boldsymbol{c}_j) \log p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta) - Q(\boldsymbol{c}_j) \log Q(\boldsymbol{c}_j) \right), \quad (5)$$

where the constant $-\sum_{i=1}^N \sum_{j=1}^K Q(\boldsymbol{c}_j) \log Q(\boldsymbol{c}_j)$ can be ignored and we need to maximize:

$$\sum_{i=1}^N \sum_{j=1}^K Q(\boldsymbol{c}_j) \log p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta). \quad (6)$$

For **the Expectation (E)-step**, $p(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta)$ (see Eq. (4)) is estimated by the ground-truth class distribution. In our approach, the number of graph prototypes (K) is identical to the number of different classes, and we generate skeleton graph prototypes $\{\boldsymbol{c}_j\}_{j=1}^K$ by computing the feature *centroids* of *encoded* skeleton graph representations $\boldsymbol{v}_i$ in different classes. We use $\{\boldsymbol{C}_j\}_{j=1}^K$ to denote the groups[2] of graph representations (referred to as *"sample groups"*) belonging to different prototypes. Then, we compute $p(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta) = \mathbb{1}(\boldsymbol{x}_i \in \boldsymbol{C}_j)$, where $\mathbb{1}(\boldsymbol{x}_i \in \boldsymbol{C}_j) = 1$ if $\boldsymbol{x}_i$ belongs to the j^{th} sample group $\boldsymbol{C}_j$ ($i.e.$, corresponding to graph prototype $\boldsymbol{c}_j$); otherwise $\mathbb{1}(\boldsymbol{x}_i \in \boldsymbol{C}_j) = 0$.

Assumption 1 ***Prototype-Class Consistency.*** *The global distribution of graph prototypes is consistent with the distribution of class feature centroids, i.e., all samples belonging to a ground-truth class explicitly correspond to the sample group of a certain prototype. In the E-step, we adopt this assumption to generate skeleton graph prototypes and derive* $p(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta) = \mathbb{1}(\boldsymbol{x}_i \in \boldsymbol{C}_j)$.

In the **Maximization (M)-step**, we combine Eq. (4) to maximize the lower-bound in Eq. (6) after the E-step:

$$\begin{aligned} &\sum_{i=1}^N \sum_{j=1}^K Q(\boldsymbol{c}_j) \log p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta) \\ &= \sum_{i=1}^N \sum_{j=1}^K p(\boldsymbol{c}_j; \boldsymbol{x}_i, \theta) \log p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta) \\ &= \sum_{i=1}^N \sum_{j=1}^K \mathbb{1}(\boldsymbol{x}_i \in \boldsymbol{C}_j) \log p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta). \end{aligned} \quad (7)$$

Assumed that each class is equally important and the sample number of each class is approximately identical in learning, each graph prototype $\boldsymbol{c}_j$ can have a uniform prior probability $p(\boldsymbol{c}_j; \theta) = \frac{1}{K}$. We have:

$$\begin{aligned} p(\boldsymbol{x}_i, \boldsymbol{c}_j; \theta) &= p(\boldsymbol{x}_i; \boldsymbol{c}_j, \theta) p(\boldsymbol{c}_j; \theta) \\ &= \frac{1}{K} \cdot p(\boldsymbol{x}_i; \boldsymbol{c}_j, \theta), \end{aligned} \quad (8)$$

where the distribution of samples around each graph prototype is assumed to be an isotropic Gaussian, leading to:

$$p(\boldsymbol{x}_i; \boldsymbol{c}_j, \theta) = \frac{\exp\left(\frac{-(\boldsymbol{v}_i - \boldsymbol{c}_p)^2}{2\sigma_p^2}\right)}{\sum_{j=1}^K \exp\left(\frac{-(\boldsymbol{v}_i - \boldsymbol{c}_j)^2}{2\sigma_j^2}\right)}, \quad (9)$$

where $\boldsymbol{v}_i = f_\theta(\boldsymbol{x}_i)$ and $\boldsymbol{c}_p$ is the graph prototype for the sample group $\boldsymbol{C}_p$ containing $\boldsymbol{x}_i$, $i.e.$, $\boldsymbol{x}_i \in \boldsymbol{C}_p$. We apply ℓ_2-normalization to both $\boldsymbol{v}$ and $\boldsymbol{c}$ to have $(\boldsymbol{v} - \boldsymbol{c})^2 = 2 - 2\boldsymbol{v} \cdot \boldsymbol{c}$. Then combining this with Eq. (2), (3), (6), (7), (8), and (9), we can get the maximum log-likelihood estimation with:

$$\begin{aligned} \theta^* &= \arg\min_\theta \sum_{i=1}^N -\log \frac{\exp(\boldsymbol{v}_i \cdot \boldsymbol{c}_p / \tau_p)}{\sum_{j=1}^K \exp(\boldsymbol{v}_i \cdot \boldsymbol{c}_j / \tau_j)} \iff \\ \theta^* &= \arg\min_\theta \sum_{k=1}^K \sum_{i=1}^{N_k} -\log \frac{\exp\left(\boldsymbol{v}_i^k \cdot \boldsymbol{c}_k / \tau_k\right)}{\sum_{j=1}^K \exp\left(\boldsymbol{v}_i^k \cdot \boldsymbol{c}_j / \tau_j\right)}, \end{aligned} \quad (10)$$

where $\boldsymbol{v}_i^k$ denotes the encoded features of i^{th} sample ($i.e.$, skeleton graph representation) belonging to the k^{th} graph prototype $\boldsymbol{c}_k$, N_k is the number of samples in the k^{th} sample group, and τ is related to the distribution of encoded graph representations around different graph prototypes.

Assumption 2 ***Maximum Homogeneous Similarity.*** *The homogeneous samples, which are defined as samples within*

[2]The graph prototypes $\{\boldsymbol{c}_j\}_{j=1}^K$ for $\{\boldsymbol{x}_i\}_{i=1}^N$ are generated based on their encoded features $\{\boldsymbol{v}_i\}_{i=1}^N$, while $\{\boldsymbol{C}_j\}_{j=1}^K$ are groups of $\{\boldsymbol{x}_i\}_{i=1}^N$ belonging to different graph prototypes.

the same-prototype sample group, should share higher inherent similarity than heterogeneous samples between different groups. In other words, the graph prototype of each sample group can represent the unique skeleton concepts and attributes of a certain identity, and the same group's samples possess the homogeneity of features corresponding to this prototype [3]. According to Assumption 1, it can be equivalent to the objective that the representation of each sample should be maximally similar to the corresponding prototype and be minimally similar to other prototypes. In the M-step, we maximize the probability that each sample representation belongs to its unique prototype (see Eq. (9)) based on this assumption. The equivalent formulation of this objective in Eq. (10) after applying feature ℓ_2-normalization can be further interpreted as to maximize the dot-product based similarity between samples and their prototypes while maximizing the dissimilarity to other prototypes.

Relations to Existing Contrastive Losses [1, 2, 4–7]:

1. The InfoNCE loss [4] re-formulated in MoCo [5] and SimCLR [6] can be interpreted as special cases of the maximum log-likelihood estimation in Eq. (10), where the prototype $\boldsymbol{c}_p$ for a feature $\boldsymbol{v}_i$ is replaced by the augmented feature $\boldsymbol{v'}_i$ generated from different views of augmentation of the same instance ($i.e.$, $\boldsymbol{c}_p = \boldsymbol{v'}_i$) and τ is fixed as a temperature for contrastive learning.

2. The masked prototype contrastive (MPC) loss in [1] and skeleton prototype contrastive (SPC) loss in [2] can be viewed as *unsupervised* generalized versions of the objective in Eq. (10). Both of them leverage unsupervised clustering algorithms ($e.g.$, DBSCAN [8]) to generate *class-agnostic* prototypes $\boldsymbol{c}_p$. The SPC loss exploits the multi-scale graph features of a skeleton sequence as $\boldsymbol{v}_i$, while the MPC loss replaces it with the features of random skeleton subsequences of a sequence. However, the instability ($e.g.$, varying cluster numbers caused by over-clustering) or/and unreliability of the used *identity-agnostic* prototypes ($e.g.$, lower confidence to characterize a ground-truth identity) largely limit their practical performance.

3. The ProtoNCE loss used in PCL [7] is a combination of momentum-based contrastive (MoCo) learning [5] and unsupervised prototype estimation with k-means clustering. It has a similar form as Eq. (10), where τ is estimated with the assumption that the distribution of feature representations around each prototype varies in different clusters. However, PCL estimates the feature distribution under the Euclidean distance metric used in the k-means clustering. Such estimation could be inapplicable ($e.g.$, can not be generalized) to models that employ different clustering algorithms ($e.g.$, density-based DBSCAN [8]) or/and different distance metrics ($e.g.$, Jaccard metric), thus failing to getting satisfactory performance in practice [1].

Temperatures. In our work, we adopt a generic form following the common practice [5,6,9], $i.e.$, setting a global temperature τ for the proposed approach. By assuming a uniform feature distribution around each instance ($i.e.$, $\tau = \tau_k = \tau_j$), we encourage the model to learn representations with higher global uniformity, which could improve the quality of contrastive representation learning as theoretically and empirically proved in [1, 10, 11].

In the proposed approach, each ground-truth identity is represented with a unique graph prototype, which is generated by computing the class centroid of encoded graph representations. We combine both *sequence-level* and *skeleton-level* graph representations to perform the graph prototype contrastive (GPC) learning, so as to learn more identity-associated graph semantics from different levels. The proposed sequence-level ($\mathcal{L}^{seq}_{\text{GPC}}$) and skeleton-level GPC loss ($\mathcal{L}^{ske}_{\text{GPC}}$) can be formulated based on Eq. (10) as:

$$\mathcal{L}^{seq}_{\text{GPC}} = \sum_{k=1}^{K}\sum_{i=1}^{N_k} -\log \frac{\exp\left(\boldsymbol{v}_i^k \cdot \boldsymbol{c}_k/\tau_1\right)}{\sum_{j=1}^{K}\exp\left(\boldsymbol{v}_i^k \cdot \boldsymbol{c}_j/\tau_1\right)}, \tag{11}$$

$$\mathcal{L}^{ske}_{\text{GPC}} = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{t=1}^{F} -\log \frac{\exp\left(\hat{\boldsymbol{v}}_{i,t}^k \cdot \hat{\boldsymbol{c}}_k/\tau_2\right)}{\sum_{j=1}^{K}\exp\left(\hat{\boldsymbol{v}}_{i,t}^k \cdot \hat{\boldsymbol{c}}_j/\tau_2\right)}, \tag{12}$$

where $\boldsymbol{c}_j$, $\hat{\boldsymbol{c}}_j$ denotes the j^{th} graph prototype and its linear projection, $i.e.$, $\hat{\boldsymbol{c}}_j = \mathcal{F}(\boldsymbol{c}_j)$, $\boldsymbol{v}_i^k$ (equivalent to $\boldsymbol{S}_{k,j}$ in the paper) denotes the graph representation of the i^{th} skeleton sequence belonging to the k^{th} class, $\hat{\boldsymbol{v}}_{i,t}^k$ (equivalent to $\mathcal{F}_1(\boldsymbol{s}_{k,j}^t)$ in the paper) represents the linear projection of graph representation of the t^{th} skeleton in the i^{th} sequence belonging to the k^{th} identity, τ_1 and τ_2 represent the global temperatures for sequence-level and skeleton-level contrastive learning, and $\mathcal{F}_1(\cdot)$, $\mathcal{F}(\cdot)$ are linear projection heads to transform skeleton-level graph representations and graph prototypes into the same contrastive space $\mathbb{R}^d$. It is worth noting that the graph prototypes are generated from higher level ($i.e.$, sequence-level) representations and the learnable linear projection in Eq. (11) can be viewed as integrating related graph features from both levels for contrastive learning. Overall, the proposed GPC loss can be viewed as a generalization of existing skeleton prototype contrastive losses [1, 2] with: **(1)** Full skeletal relation encoding, which exploits Skeleton Graph Transformer (SGT) to *simultaneously* captures structural and actional relations from both adjacent and non-adjacent body joints (see Sec. 3.2 of the paper); **(2)** Fine-grained skeleton semantics learning, which combines graph prototype contrastive learning

(sequence and skeleton level semantics, see Sec. 3.3 of the paper) and graph structure-trajectory prompted reconstruction (graph and node level semantics, see Sec. 3.4 of the paper).

Convergence Proof

We prove the convergence of GPC under modeling the maximum log-likelihood estimation (see Eq. (10)). Recall Eq. (2) and (3) and let

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{N} \log p\left(\boldsymbol{x}_i;\theta\right) \\ &= \sum_{i=1}^{N} \log \sum_{j=1}^{K} p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right) \\ &= \sum_{i=1}^{N} \log \sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right) \frac{p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)}{Q\left(\boldsymbol{c}_j\right)} \\ &\geq \sum_{i=1}^{N} \sum_{j=1}^{K} Q\left(\boldsymbol{c}_j\right) \log \frac{p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta\right)}{Q\left(\boldsymbol{c}_j\right)}. \end{aligned} \tag{13}$$

The above inequality holds with equality when $Q\left(\boldsymbol{c}_j\right) = p\left(\boldsymbol{c}_j;\boldsymbol{x}_i,\theta\right)$ is a constant (see Eq. (4)).

In the t^{th} E-step, we have estimated the constant value $Q^{(t)}\left(\boldsymbol{c}_j\right) = p\left(\boldsymbol{c}_j;\boldsymbol{x}_i,\theta^{(t)}\right)$ based on the ground-truth class distribution. Then we have:

$$\ell\left(\theta^{(t)}\right) = \sum_{i=1}^{N} \sum_{j=1}^{K} Q^{(t)}\left(\boldsymbol{c}_j\right) \log \frac{p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta^{(t)}\right)}{Q^{(t)}\left(\boldsymbol{c}_j\right)}. \tag{14}$$

For the t^{th} M-step, we fix $Q^{(t)}\left(\boldsymbol{c}_j\right) = p\left(\boldsymbol{c}_j;\boldsymbol{x}_i,\theta^{(t)}\right)$ and train model parameters θ to maximize Eq. (14). In this way, we can always have:

$$\begin{aligned}\ell\left(\theta^{(t+1)}\right) &\geq \sum_{i=1}^{N} \sum_{j=1}^{K} Q^{(t)}\left(\boldsymbol{c}_j\right) \log \frac{p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta^{(t+1)}\right)}{Q^{(t)}\left(\boldsymbol{c}_j\right)} \\ &\geq \sum_{i=1}^{N} \sum_{j=1}^{K} Q^{(t)}\left(\boldsymbol{c}_j\right) \log \frac{p\left(\boldsymbol{x}_i,\boldsymbol{c}_j;\theta^{(t)}\right)}{Q^{(t)}\left(\boldsymbol{c}_j\right)} \\ &= \ell\left(\theta^{(t)}\right). \end{aligned} \tag{15}$$

The above result that $\ell\left(\theta^{(t)}\right)$ monotonically increases with more iterations suggests the convergence of the algorithm.

The detailed convergence properties of the EM algorithm are discussed in [12, 13]. Here we only discuss the general case, and follow [13] to make the assumptions for the EM algorithm:

- **(a)** Ω is a subset in the r-dimensional Euclidean space $\mathbb{R}^r$.
- **(b)** $\Omega_{\theta^{(0)}} = \left\{\theta \in \Omega : \ell(\theta) \geq \ell\left(\theta^{(0)}\right)\right\}$ is compact for any $\ell\left(\theta^{(0)}\right) > -\infty$.
- **(c)** $\ell(\cdot)$ is continuous in Ω and differentiable in the interior of Ω.

Under the assumptions of **(a)**, **(b)**, and **(c)**[3], we have: **(d)** $\left\{\ell\left(\theta^{(t)}\right)\right\}_{t\geq 0}$ is bounded above for any $\theta^{(0)} \in \Omega$. As a consequence of **(d)** and the inequality (15) (*i.e.*, $\ell\left(\theta^{(t+1)}\right) \geq \ell\left(\theta^{(t)}\right)$), $\ell\left(\theta^{(t)}\right)$ converges monotonically to some ℓ^*.

It is worth noting that there is no guarantee that ℓ^* is the global maximum of $\ell(\cdot)$ over Ω. As reported in previous works [13–17], if the log-likelihood function $\ell(\cdot)$ has several (local or global) maxima and stationary points, the convergence of the EM sequence $\left\{\ell\left(\theta^{(t)}\right)\right\}$ to either type of point depends on the choice of starting point. Readers can refer to [12, 13] for more details about different convergence cases of the EM algorithm.

The aforementioned fact may account for the performance changes (*i.e.*, small variations) of our model on the same dataset, as different random initializations of model parameters could change $\theta^{(0)}$ (*i.e.*, the starting point) hence the final convergence result. In practice, we follow [1, 18] to train the model with different random initializations on each dataset and report its average performance, which helps estimate a more stable EM convergence result with different initialized starting points.

References

[1] H. Rao and C. Miao, "SimMC: Simple masked contrastive learning of skeleton representations for unsupervised person re-identification," in *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1290–1297, 2022. 2, 3, 4

[2] H. Rao and C. Miao, "Skeleton prototype contrastive learning with multi-level graph relation modeling for unsupervised person re-identification," *arXiv preprint arXiv:2208.11814*, 2022. 2, 3

[3] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI," *arXiv preprint arXiv:2201.08164*, 2022. 3

[4] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. 3

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020. 3

[3] These assumptions can be satisfied in most practical situations. As the related proofs/discussions are out of the scope of this work, readers can refer to [13] for more details.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020. 3

[7] J. Li, P. Zhou, C. Xiong, and S. Hoi, "Prototypical contrastive learning of unsupervised representations," in *International Conference on Learning Representation (ICLR)*, 2021. 3

[8] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, vol. 96, pp. 226–231, 1996. 3

[9] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3733–3742, 2018. 3

[10] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International Conference on Machine Learning (ICML)*, pp. 9929–9939, 2020. 3

[11] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6894–6910, 2021. 3

[12] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007. 4

[13] C. J. Wu, "On the convergence properties of the EM algorithm," *The Annals of statistics*, pp. 95–103, 1983. 4

[14] V. Hasselblad, "Estimation of finite mixtures of distributions from the exponential family," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1459–1471, 1969. 4

[15] J. H. Wolfe, "Pattern clustering by multivariate mixture analysis," *Multivariate behavioral research*, vol. 5, no. 3, pp. 329–350, 1970. 4

[16] N. Laird, "Nonparametric maximum likelihood estimation of a mixing distribution," *Journal of the American Statistical Association*, vol. 73, no. 364, pp. 805–811, 1978. 4

[17] D. B. Rubin and D. T. Thayer, "EM algorithms for ml factor analysis," *Psychometrika*, vol. 47, no. 1, pp. 69–76, 1982. 4

[18] H. Rao, S. Wang, X. Hu, M. Tan, Y. Guo, J. Cheng, X. Liu, and B. Hu, "A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 01, pp. 1–1, 2021. 4