Skeleton Prototype Contrastive Learning with Multi-Level Graph Relation Modeling for Unsupervised Person Re-Identification

Haocong Rao, Chunyan Miao

Abstract—Person re-identification (re-ID) via 3D skeletons is an important emerging topic with many merits. Existing solutions rarely explore valuable body-component relations in skeletal structure or motion, and they typically lack the ability to learn general representations with unlabeled skeleton data for person re-ID. This paper proposes a generic unsupervised Skeleton Prototype Contrastive learning paradigm with Multi-level Graph Relation learning (SPC-MGR) to learn effective representations from unlabeled skeletons to perform person re-ID. Specifically, we first construct unified multi-level skeleton graphs to fully model body structure within skeletons. Then we propose a multi-head structural relation layer to comprehensively capture relations of physically-connected body-component nodes in graphs. A full-level collaborative relation layer is exploited to infer collaboration between motion-related body parts at various levels, so as to capture rich body features and recognizable walking patterns. Lastly, we propose a skeleton prototype contrastive learning scheme that clusters feature-correlative instances of unlabeled graph representations and contrasts their inherent similarity with representative skeleton features (“skeleton prototypes”) to learn discriminative skeleton representations for person re-ID. Empirical evaluations show that SPC-MGR significantly outperforms several state-of-the-art skeleton-based methods, and it also achieves highly competitive person re-ID performance for more general scenarios.

Index Terms—Skeleton Based Person Re-Identification; Unsupervised Representation Learning; Multi-Level Skeleton Graphs; Skeleton Prototype Contrastive Learning

1 INTRODUCTION

PERSON re-identification (re-ID) aims at identifying or matching a target pedestrian across different views or scenes, which plays an essential role in safety-critical applications including intelligent video surveillance, security authentication and human tracking [1]–[9]. Conventional studies [10]–[16] typically utilize visual features such as human appearances, silhouettes and body textures from RGB or depth images to discriminate different individuals. Nevertheless, this kind of methods are often vulnerable to appearance, lighting and clothing variation in practice. Compared with RGB-based and depth-based methods, 3D skeleton-based models [17]–[22] exploit 3D coordinates of numerous key joints to characterize human body and motion, which enjoys smaller data size and better robustness to scale and view variation [23]. With these advantages, 3D skeleton data have drawn surging attention in the fields of person re-ID and gait recognition [19]–[22], [24], [25]. However, the way to model discriminative body and motion features with 3D skeleton data remains to be an open challenge.

To perform person re-ID via 3D skeletons, existing endeavors typically model skeleton features by two groups of methods. Skeleton descriptor based methods [17], [18], [26] manually extract certain anthropometric and geometric attributes of body from skeleton data. However, these hand-crafted methods usually require domain knowledge such as human anatomy [27], and cannot fully mine underlying features beyond human cognition. Deep neural network based methods [19], [20], [25] usually leverage convolutional neural networks (CNN) or long short-term memory (LSTM) to learn skeleton representations with sequences of raw body-joint positions or pose descriptors (e.g., limb lengths). Nevertheless, these works rarely explore inherent relations between different body joints or components, which could ignore some valuable structural information of human body. Taking the human walking for example, neighbor body joints “foot” and “knee” have strong motion correlations, while they usually enjoy diverse degree of collaboration with limb-level components “leg” and “arm” during movement, which can be exploited to catch unique and recognizable patterns [28]. Other important flaws of this type of methods are label dependency and weak generalization ability. In
practical terms, these methods usually require massive labeled data of pre-defined classes to either train the model from scratch [21], [25] or fine-tune the pre-trained skeleton representations [19], [20], [22] to classify the known identities. As a result, they lack the flexibility to learn general and representative skeleton features that can re-identify different pedestrians under the unavailability of labels, which limits its application in many real-world scenarios.

To address the above challenges, this work for the first time proposes a generic Skeleton Prototype Contrastive learning paradigm with Multi-level Graph Relation modeling (SPC-MGR) in Fig. 1 that can comprehensively model body structure and relations at various levels and mine discriminative features from unlabeled skeletons for person re-ID. Specifically, we first devise multi-level graphs to represent each 3D skeleton in a unified coarse-to-fine manner, so as to fully model body structure within skeletons. Then, to enable a comprehensive exploration of relations between different body components, we propose to model structural-collaborative body relations within skeletons from multi-level graphs. In particular, since each body component is highly correlated with its physically-connected components and may possess different structural relations (e.g., motion correlations), we propose a multi-head structural relation layer (MSRL) to capture multiple relations between each body-component node and its neighbors within a graph, so as to aggregate key correlative features for effective node representations. Meanwhile, motivated by the fact that dynamic cooperation of body components in motion could carry unique patterns (e.g., gait) [28], we propose a full-level collaborative relation layer (FCRL) to adaptively infer collaborative relations among motion-related components at both the same-level and cross-level in graphs. Furthermore, we exploit a multi-level graph feature fusion strategy to integrate features of different-level graphs via collaborative relations, which encourages the model to capture more graph structural semantics and discriminative skeleton features. Lastly, to mine effective features from unlabeled skeleton graph representations (referred as skeleton instances), we propose a skeleton prototype contrastive learning scheme (SPC), which clusters correlative skeleton instances and contrasts their inherent similarity with the most representative skeleton features (referred as skeleton prototypes) to learn general discriminative skeleton representations in an unsupervised manner. By maximizing the similarity of skeleton instances to their corresponding prototypes and their dissimilarity to other prototypes, SPC encourages the model to capture more discriminative skeleton features and class-related semantics (e.g., intra-class similarity) for person re-ID without using any label.

An earlier and preliminary version of this work was reported in [21]. Compared with [21] that performs supervised skeleton representation for person re-ID, this paper focuses on the unsupervised re-ID problem under general settings. In this work, we explore unified and generalized multi-level skeleton graphs to extend earlier graph representations to different skeleton data, and propose a novel skeleton prototype contrastive learning paradigm to mine discriminative skeleton features with the proposed unlabeled skeleton graphs for person re-ID. To our best knowledge, this is also the first work that explores a generic unsupervised learning paradigm for skeleton-based person re-ID. We also devise a new full-level collaborative relation layer to capture more comprehensive relations between different-level body components, and further explore a multi-level graph feature fusion strategy to enhance graph semantics and global pattern learning. To validate the above improvement, this work conducts comprehensive experiments on four public person re-ID datasets, and carries out more detailed analysis on different person re-ID tasks (e.g., multi-view person re-ID) under more general scenarios. Furthermore, we propose two evaluation metrics, mean intra-Class Tightness (mACT) and mean inter-Class Looseness (mRCL), to quantitatively evaluate the tightness of same-class representations and the looseness of different-class representations, so as to showcase the effectiveness of skeleton representations learned by the proposed approach.

Our main contributions are summarized as follows:

- We devise unified multi-level graphs to model 3D skeletons, and propose a novel Skeleton Prototype Contrastive learning paradigm with Multi-level Graph Relation modeling (SPC-MGR) to learn an effective representation from unlabeled skeleton data for unsupervised person re-ID.
- We propose multi-head structural relation layer (MSRL) to capture relations of neighbor body components, and devise full-level collaborative relation layer (FCRL) to infer collaboration between different-level components, so as to learn more structural semantics and unique patterns.
- We propose a skeleton prototype contrastive learning (SPC) scheme to encourage capturing representative discriminative skeleton features and high-level class-related semantics from unlabeled skeleton data for person re-ID.
- We propose mean intra-class tightness (mACT) and mean inter-class looseness (mRCL) to evaluate skeleton representation learning, and show the effectiveness of skeleton representations learned by our unsupervised approach.
- Extreme experiments show that the proposed SPC-MGR outperforms several state-of-the-art skeleton-based methods on four person re-ID benchmarks, and is also highly effective when applied to skeleton data estimated from large-scale RGB videos under more general re-ID settings.

The rest of our paper is organized as follows. Sec. 2 provides an introduction for related works in the literature. Sec. 3 presents each technical module of the proposed approach. Sec. 4 elucidates the details of experiments, and comprehensively compares our approach with existing solutions. Sec. 5 conducts ablation study and extensive analysis of the proposed approach. Sec. 6 concludes this paper. Sec. 7 provides a statement for potential ethical issues.

2 RELATED WORKS

2.1 Skeleton-Based Person Re-Identification

2.1.1 Hand-Crafted Methods

Early skeleton-based works extract hand-crafted descriptors in terms of certain geometric, morphological or anthropometric attributes of human body. Barbosa et al. [17] compute 7 Euclidean distances between the floor plane and joint or joint pairs to construct a distance matrix, which is learned by a quasi-exhaustive strategy to extract discriminative features for person re-ID. Munaro et al. [26] and Pala et al. [29] further extend them to 13 ($D_{13}$) and 16 skeleton descriptors ($D_{16}$) respectively, and leverage support vector machine (SVM), k-nearest neighbor (KNN) or Adaboost classifiers for person re-ID. Since such solutions using 3D skeletons alone are hard to achieve satisfactory performance, they usually combine other modalities such as 3D point clouds [30] and 3D face descriptors [29] to improve person re-ID accuracy.
2.1.2 Supervised and Self-Supervised Methods

Most recently, a few works exploit deep learning paradigms to learn gait representations from skeleton data for person re-ID in a supervised or self-supervised manner. Liao et al. [25] propose PoseGait, which feeds 81 hand-crafted pose features of 3D skeletons into CNNs for human recognition. Rao et al. [19] devise a self-supervised attention-based gait encoding (AGE) model with multi-layer LSTM to encode gait features from unlabeled skeleton sequences, and then fine-tune the learned features with the supervision of labels for person re-ID. In [20], they further propose a locality-awareness approach (SGELA) that combines various pretext tasks (e.g., reverse sequential reconstruction) and contrastive learning scheme to enhance self-supervised gait representation learning for the person re-ID task. The latest self-supervised work is SM-SGE [22], which utilizes a skeleton graph based reconstruction and inference mechanism to encode discriminative skeleton structure and motion features for the person re-ID task.

Our work differs in following aspects. We for the first time propose an unsupervised learning paradigm to learn discriminative skeleton representations from unlabeled skeleton data to directly perform person re-ID. Our approach requires neither any form of feature engineering (e.g., hand-crafted skeleton descriptors [17], [26], [29]) nor any skeleton label for training model [21], [25] or fine-tuning representations [19], [20], [22]. Unlike most existing works that directly represent skeletons as sequences of body joints [19], [20] or pose descriptors [25], this work not only explores more comprehensive body-structure modeling by devising unified and generalized multi-level skeleton graphs, but also systematically mines valuable body-component relations at various levels. Besides, the proposed approach needs neither self-supervised pretext tasks (e.g., reconstruction [22]) for pre-training or supervised classification networks for discriminative representation learning. Instead, it exploits the unsupervised skeleton prototype contrastive learning to mine representative skeleton features and class-related semantics from unlabeled 3D skeleton data. To the best of our knowledge, this is also the first attempt to leverage skeleton graph based relation modeling and contrastive learning to realize both skeleton representation learning and unsupervised person re-ID.

2.2 Contrastive Learning

Contrastive learning has recently achieved great success in many self-supervised and unsupervised learning tasks [16], [20], [31]–[35]. Its general objective is to learn effective data representations by pulling closer positive pairs and pushing apart negative pairs in the feature space using contrastive losses, which are often designed based on certain auxiliary tasks (e.g., similarity metrics learning). For example, Wu et al. [32] devise an instance-level discrimination method in the form of exemplar task [36] to perform image contrastive learning with noise-contrastive estimation loss (NCE) [37]. In [33], contrastive predictive coding (CPC) based on a probabilistic contrastive loss (InfoNCE) is proposed to learn general representations for different domains. To optimize representation learning (e.g., consistency) in memory bank based contrastive methods [38]–[40], some recent end-to-end works [41]–[43] utilize all samples of the current mini-batch to generate negative instance features, while the momentum-based approach [44] further explores the use of momentum-updated encoder and queue dictionary to improve consistency of both encoder and instance features. The latest PCL [35] integrates both contrastive learning and clustering into an expectation-maximization (EM) framework, which is highly efficient on unsupervised visual representation learning and inspires our work for 3D skeletons.

Our work differs from previous studies in following aspects. The proposed skeleton prototype contrastive learning scheme is devised to mine most representative and discriminative skeleton features (i.e., skeleton prototypes) from unlabeled multi-level skeleton graph representations of 3D skeleton sequences (i.e., skeleton instances). In this scheme, the skeleton instances and generated skeleton prototypes are exploited as the contrasting instances, which fundamentally differs from existing works [34], [35], [43], [44] that use augmented samples of images as instances. The goal of skeleton prototype contrastive learning scheme is to maximize skeleton instances’ similarity with their prototypes and dissimilarity to other prototypes using the proposed clustering-contrastive strategy, which does not require any extra memory or sampling mechanism [16], [35] and can directly learn effective skeleton representations for the person re-ID task without introducing any ground-truth label from source data domains.

3 The Proposed Approach

Suppose that a 3D skeleton sequence \( S_{1:T} = \{ S_1, \ldots, S_T \} \subset \mathbb{R}^{J \times X \times D} \) where \( S_t \subset \mathbb{R}^{J \times X \times D} \) is the \( t \)th skeleton with \( J \) body joints and \( D = 3 \) dimensions. Each skeleton sequence \( S_{1:T} \) corresponds to an ID label \( y \), where \( y \in \{ 1, \ldots, C \} \) and \( C \) is the number of different persons. The training set \( \Phi_1 = \{ S_{1:f}^{n_l} \}_{l=1}^{N_l} \) probe set \( \Phi_p = \{ S_{1:f}^{n_l} \}_{l=1}^{N_2} \) and gallery set \( \Phi_g = \{ S_{1:f}^{n_l} \}_{l=1}^{N_3} \) contain \( N_1 \), \( N_2 \), and \( N_3 \) skeleton sequences of different persons under varying views or scenes. Our goal is to learn an embedding function \( \psi(\cdot) \) that maps \( \Phi_p \) and \( \Phi_g \) to effective skeleton representations \( \{ \mathbf{M}_i^{p,:} \}_{i=1}^{N_2} \) and \( \{ \mathbf{M}_j^{g,:} \}_{j=1}^{N_3} \) without using any label, such that the representation \( \mathbf{M}_i^{p,:} \) in the probe set can match the representation \( \mathbf{M}_j^{g,:} \) of the same identity in the gallery set. The overview of the proposed approach is given in Fig. 2, and we present the details of each technical component below.

3.1 Unified Multi-Level Skeleton Graphs

Inspired by the fact that human motion can be decomposed into movements of functional body-components (e.g., legs, arms) [45], we spatially group skeleton joints to be higher level body components at their centroids. Specifically, we first divide human skeletons into several partitions from coarse to fine. Based on the nature of body structure, we specify the location of each body partition and its corresponding skeleton joints of different sources (e.g., datasets). Then, we adopt the weighted average of body joints in the same partition as the node of higher level body component and use its physical connections as edges, so as to build unified skeleton graphs for an input skeleton. As shown in Fig. 3, we construct three levels of skeleton graphs, namely part-level, body-level and hyper-body-level graph for each skeleton \( \mathbf{S} \), which can be represented as \( G^1 \), \( G^2 \) and \( G^3 \) respectively. Each graph \( G^l (\mathbf{V}^l, \mathbf{E}^l) \) (\( l \in \{ 1, 2, 3 \} \)) consists of nodes \( \mathbf{V}^l = \{ \mathbf{v}_1^l, \mathbf{v}_2^l, \ldots, \mathbf{v}_n^l \} \), \( \mathbf{E}^l \subset \mathbb{R}^{D_l} \) with \( l \in \{ 1, \ldots, n_l \} \) and edges \( \mathbf{E}^l = \{ \mathbf{e}_{i,j}^l \mid \mathbf{v}_i^l, \mathbf{v}_j^l \in \mathbf{V}^l \}, \mathbf{e}_{i,j}^l \in \mathbb{R} \). Here \( \mathbf{V}^l \) and \( \mathbf{E}^l \) denote the set of nodes corresponding to different body components and the set of their internal connection relations, respectively. \( n_l \) denotes the number of nodes in \( G^l \). More formally, we define a graph’s adjacency matrix as \( \mathbf{A}_l \in \mathbb{R}^{n_l \times n_l} \) to represent structural relations among \( n_l \) nodes. We compute the normalized structural
Spatially Group Part-Level Graph Body-Level Graph Hyper-Body-Level Graph Averagely Merge

Multi-Level Skeleton Graphs

Fig. 2. Schematic diagram of SPC-MGR. Firstly, each 3D skeleton of an input sequence $S_1, S_2, \cdots, S_f$ is represented with part-level, body-level, and hyper-body-level graphs. Secondly, we employ multi-head structural relation layers (MSRL) to capture structural relations of neighbor nodes in each graph, and averagely aggregate features learned by multiple heads to obtain node representations. Then, full-level collaborative relation layers (FCRL) infer the dynamic collaborative relations among the same-level and different-level body components, which are exploited to integrate key graph features into multi-level skeleton graph representations $F^1$, $F^2$, and $F^3$. Next, we perform clustering on skeleton instances, which are sequence-level (“Seq.”) multi-level skeleton graph representations, to generate clusters and corresponding skeleton prototypes. Finally, during skeleton prototype contrastive learning, we enhance the similarity of instances belonging to the same prototype and maximize their dissimilarity to other prototypes by minimizing contrastive loss $\mathcal{L}_{SPC}$. The learned skeleton graph representations are exploited to perform person re-ID.

relations between node $i$ and its neighbors, i.e., $\sum_{j \in \mathcal{N}_i} A_{i,j} = 1$, where $\mathcal{N}_i$ denotes the neighbor nodes of node $i$. $A_i$ is adaptively learned to capture flexible structural relations in the training stage.

Remarks: The proposed unified multi-level skeleton graphs can be generalized into different skeleton datasets, which enables the pre-trained model to be directly transferred across different domains for generalized person re-ID (see Sec. 5.5.2). Besides, they can be extended to skeleton data estimated from RGB videos to learn effective person re-ID representations (see Sec. 5.5.1).

3.2 Structural-Collaborative Body Relation Modeling

The physical connections of body structure typically endow body components in a local partition with higher correlations, while components of different parts may act collaboratively in various global patterns during motion [46]. To exploit such internal relations to mine rich body-structure features and unique motion characteristics from skeletons, we propose the multi-level structural relation layer (MSRL) and full-level collaborative relation layer (FCRL) to model the structural and collaborative relations of body components from multi-level skeleton graphs as follows.

3.2.1 Multi-Head Structural Relation Layer

To capture latent body structural information and learn an effective representation for each body-component node in skeleton graphs, we propose to focus on features of structurally-connected neighbor nodes, which enjoy higher correlations (referred as structural relations) than distant pairs. For instance, adjacent nodes usually have closer spatial positions and similar motion tendency. Therefore, we devise a multi-head structural relation layer (MSRL) to learn relations of neighbor nodes and aggregate the most correlative spatial features to represent each body-component node.

We first devise a basic structural relation head based on the graph attention mechanism [47], which can focus on more correlative neighbor nodes by assigning larger attention weights, to capture the internal relation $e_{i,j}$ between adjacent nodes $i$ and $j$ in the same graph as:

$$e_{i,j} = \text{LeakyReLU} \left( W^e_{i,j} \left[ W_{u_i} u_i'^T \right] \left[ W_{u_j} u_j'^T \right] \right)$$

(1)
where $\mathbf{W}_{l}^i \in \mathbb{R}^{D \times D_h}$ denotes the weight matrix to map the $l^{th}$ level node features $\mathbf{v}_i^l \in \mathbb{R}^{D}$ into a higher level feature space $\mathbb{R}^{D_h}$, $\mathbf{W}_{l}^i \in \mathbb{R}^{2D_h}$ is a learnable weight matrix to perform relation learning in the $l^{th}$ level graph, $\|$ indicates concatenating features of two nodes, and LeakyReLU($\cdot$) is a non-linear activation function. Then, to learn flexible structural relations to focus on more correlative nodes, we normalize relations using the softmax function as follows:

$$A_{l,i,j} = \text{softmax}_j \left( \frac{\exp(c_{i,j})}{\sum_{k \in N_i} \exp(c_{i,k})} \right)$$

where $N_i$ denotes directly-connected neighbor nodes (including $i$) of node $i$ in graph. We use structural relations $A_{l,i,j}^{\theta}$ to aggregate features of most relevant nodes to represent node $i$:

$$\tilde{\mathbf{v}}_i^l = \sigma \left( \sum_{j \in N_i} A_{l,i,j}^{\theta} \mathbf{W}_{l}^i \mathbf{v}_j^l \right)$$

where $\sigma(\cdot)$ is a non-linear function and $\tilde{\mathbf{v}}_i^l \in \mathbb{R}^{D_h}$ is feature representation of node $i$ computed by a structural relation head.

To sufficiently capture potential structural relations (e.g., position similarity and movement correlations) between each node and its neighbor nodes, we employ multiple structural relation heads, each of which independently executes the same computation of Eq. 3 to learn a potentially different structural relation, as shown in Fig. 4. We averagely aggregate features learned by $m$ different structural relation heads as the representation of node $i$ as follows:

$$\bar{\mathbf{v}}_i^l = \frac{1}{m} \sum_{s=1}^{m} \sigma \left( \sum_{j \in N_i} (A_{l,i,j}^{\theta})^s (\mathbf{W}_{l}^i)^s \mathbf{v}_j^l \right)$$

where $\bar{\mathbf{v}}_i^l \in \mathbb{R}^{D_h}$ denotes the multi-head feature representation of node $i$ in graph, $m$ is the number of structural relation heads, $(A_{l,i,j}^{\theta})^s \in \mathbb{R}$ represents the structural relation between node $i$ and $j$ computed by the $s^{th}$ structural relation head, and $(\mathbf{W}_{l}^i)^s$ denotes the corresponding weight matrix to perform feature mapping in the $s^{th}$ head. Here we use average rather than concatenation operation to reduce feature dimension and allow for more structural relation heads. MSRL enables our model to capture the relations of correlative neighbor nodes (see Eq. 1 and 2) and integrates key spatial features into node representations of each graph (see Eq. 3 and 4). However, it only considers the local relations of the same-level components in graphs and is insufficient to capture global collaboration between different level body components, which motivates us to propose the full-level collaborative relation layer.

### 3.2.2 Full-Level Collaborative Relation Layer

Motivated by the natural property of human walking, i.e., gait, which could be represented by the dynamic cooperation among body joints or between different body components [28], we expect our model to infer the degree of collaboration (referred as collaborative relations) among body-component nodes in multi-level graphs, so as to capture more unique and recognizable walking patterns from the motion of skeletons. For this purpose, we propose a full-level collaborative relation layer (FCRL) to capture relations between a node and all motion-related nodes of the same level and that between a node and its spatially corresponding higher level body component or other potential components. As shown in Fig. 2 and Fig. 4, we compute collaborative relation matrix $\hat{A}_{a,i,j}^{a,b} \in \mathbb{R}^{n_a \times n_b}$ ($a, b \in \{1, 2, 3\}, a \leq b$) between the $a^{th}$ level nodes $\mathbf{v}_i^a$ and the $b^{th}$ level nodes $\mathbf{v}_j^b$ as following:

$$\hat{A}_{a,i,j}^{a,b} = \text{softmax}_j \left( \frac{\exp(v_i^a \top \mathbf{v}_j^b)}{\sum_{k=1}^{n_b} \exp(v_i^a \top \mathbf{v}_k^b)} \right)$$

where $\hat{A}_{a,i,j}^{a,b}$ is the collaborative relation between node $i$ in graph $G^a$ and node $j$ in graph $G^b$. Here we use the inner product of multi-head node feature representations (see Eq. 4) that retain key spatial information of nodes to measure the degree of collaboration. Instead of merely considering body relations between adjacent level graphs [21], FCRL can capture the global collaborative relations among both adjacent and non-adjacent graphs, and meanwhile provides full collaboration inference between a node and all potential motion-correlated nodes in the same graph.

### 3.2.3 Multi-Level Graph Feature Fusion

To enhance structural semantics of multiple graphs (e.g., global graph patterns) and adaptively integrate key correlative features in component collaboration, we exploit collaborative relations to fuse body-component node features across different spatial levels. We update the node representation ($\tilde{v}_i^a$) of $a^{th}$ level graph by fusing collaborative node features ($\hat{\tilde{v}}_i^a$) learned from different graphs:

$$\bar{\tilde{v}}_i^a = \tilde{v}_i^a + \sum_{b=1}^{3} \left( \lambda_{C} \sum_{j=1}^{n_a} A_{a,i,j}^{a,b} \mathbf{W}_{C}^{a,b} \mathbf{v}_j^b \right)$$

where $\mathbf{W}_{C}^{a,b} \in \mathbb{R}^{D_h \times D_h}$ is a learnable weight matrix to integrate features of collaborative node $\hat{\tilde{v}}_i^b$ of $b^{th}$ level into $a^{th}$ level node $\bar{\tilde{v}}_i^a$. $n_b$ denotes the number of nodes in the $b^{th}$ level graph, and $\lambda_{C}$ represents the fusion coefficient between $a^{th}$ level and $b^{th}$ level graphs, which can be adjusted according to their inherent correlations (e.g., level similarity). We denote the fused $l^{th}$ level graph features of the $i^{th}$ skeleton as $\bar{\mathbf{F}}_i^l \in \mathbb{R}^{n_1+n_2+n_3 \times D_h}$ by concatenating all node representations. Inspired by [22], we retain graph representations of each individual level and adopt their concatenation to represent a skeleton as follows:

$$M_i = [\mathbf{F}_i^1; \mathbf{F}_i^2; \mathbf{F}_i^3]$$

where $M_i \in \mathbb{R}^{(n_1+n_2+n_3) \times D_h}$ is the multi-level graph representation of the $i^{th}$ skeleton $\mathbf{S}_i$, and $[;]$ indicates the concatenation of
graph features. By combining all graph-level representations that integrate structural and collaborative body relation features (see Eq. 1-Eq. 6), we encourage the model to capture richer features of body structure and skeleton patterns at various levels. We show the effectiveness of the proposed structural-collaborative body relation learning by the relation visualization in Sec. 5.3 and Appendix.

3.3 Skeleton Prototype Contrastive Learning Scheme

As skeletons of the same individual typically share highly similar body attributes (e.g., anthropometric attributes) and unique walking patterns [28], it is natural to consider mining the most typical attributes or patterns to identify the same person from others. To achieve this goal and encourage the model to capture more high-level skeleton semantics (e.g., class-related patterns), we propose a Skeleton Prototype Contrastive learning (SPC) scheme to focus on the most representative skeleton graph features (referred as skeleton prototypes) of pedestrians and exploit their inherent similarity and dissimilarity with other unlabeled graph representations (referred as skeleton instances) to learn general and discriminative representations of each individual for the person re-ID task.

3.3.1 Skeleton Prototype Generation

Given multi-level graph representations \( \{M_1, \cdots, M_f \} \) of an input skeleton sequence \( (S_{1:f} = (S_1, \cdots, S_f)) \), we first integrate graph features into a sequence-level skeleton graph representation:

\[
\bar{M} = \frac{1}{f} \sum_{i=1}^{f} w_i M_i \tag{8}
\]

where \( \bar{M} \) is the multi-level graph representation of skeleton sequence \( S_{1:f} \), which incorporates structural-collaborative features and temporal dynamics of \( f \) consecutive multi-level skeleton graphs, and \( w_i \) denotes the importance of \( i^{th} \) skeleton graph representation. Here we assume that each skeleton uniquely contributes to representing graph features of a sequence, i.e., \( w_i = 1 \). For clarity, we use \( \bar{M} = \{\bar{M}_i\}_{i=1}^{N_i} \) to represent multi-level graph representations of skeletons in the training set \( \mathcal{F}_t \), which are exploited as skeleton instances in the proposed SPC scheme.

Then, to gather skeleton instances \( \bar{M} \) that contain similar features to find the representative skeleton prototypes, we leverage the DBSCAN algorithm [48], which can discover clusters with arbitrary shapes or semantics, to perform clustering as:

\[
\text{DBSCAN}(\bar{M}) \rightarrow \bar{M}^1, \bar{M}^2, \cdots, \bar{M}^z, \bar{M}^o \tag{9}
\]

where \( \bar{M} = \bar{M}^1 \cup \bar{M}^2 \cup \cdots \cup \bar{M}^z \cup \bar{M}^o, z \) is the number of clusters (i.e., pseudo classes), \( \bar{M}^k = \{\bar{M}^k_i\}_{i=1}^{z_k}, k \in \{1, \cdots, z\} \), is the cluster that contains \( z_k \) instances belonging to the \( k^{th} \) pseudo class, and \( \bar{M}^o = \{\bar{M}^o_i\}_{i=1}^{z_o} \) denotes the set of outlier instances that do not belong to any cluster. We compute the centroid of each cluster, which averagely aggregates features of skeleton instances in the cluster, to obtain corresponding skeleton prototype:

\[
P^k = \frac{1}{z_k} \sum_{i=1}^{z_k} \bar{M}^k_i \tag{10}
\]

where \( \bar{M}^k_i \in \mathbb{R}^{(n_1+n_2+n_3) \times D_h} \) is the \( i^{th} \) skeleton instance in the \( k^{th} \) cluster, and \( P^k \) denotes the \( k^{th} \) skeleton prototype.

3.3.2 Skeleton Prototype Contrastive Learning

To focus on the typical and discriminative features of skeleton prototypes as well as facilitate learning high-level skeleton semantics from different prototypes, we propose to enhance the inherent similarity of a skeleton instance to corresponding skeleton prototype and maximize its dissimilarity to other skeleton prototypes with a skeleton prototype contrastive loss as:

\[
L_{\text{SPC}} = \frac{1}{N} \sum_{k=1}^{z} \sum_{i=1}^{z_k} \log \frac{\exp \left( \frac{\| \bar{M}^k_i \cdot P^k / \tau \|}{\sum_{j=1}^{z} \exp \left( \frac{\| \bar{M}^j \cdot P^j / \tau \|}{\tau} \right) } \right)}{\sum_{j=1}^{z} \exp \left( \frac{\| \bar{M}^j \cdot P^k / \tau \|}{\tau} \right) } \tag{11}
\]

where \( N \) represents the number of all training instances, \( z \) denotes the number of skeleton prototypes, \( x_k \) is the number of skeleton instances belonging to the \( k^{th} \) prototype \( P^k \), and \( \tau \) represents the temperature for contrastive learning, where higher value of \( \tau \) produces a softer probability distribution over prototypes and retains more similar information among clusters [49]. The skeleton prototype contrastive loss \( L_{\text{SPC}} \) can be viewed as a special form of InfoNCE loss [33], which aims to maximize the dot product based similarity between the query for each skeleton instance and its positive key for the same prototype while maximizing the dissimilarity to negative keys for other prototypes.

3.3.3 Analysis of Skeleton Prototype Contrastive Learning

The objective of skeleton prototype contrastive learning is to maximize the intra-prototype similarity and inter-prototype dissimilarity to mine discriminative skeleton features and identity-related semantics in an unsupervised manner. Such learning can be interpreted as to simultaneously enhance the tightness of skeleton instances belonging to the same prototype and the looseness among different prototypes. Considering that a better representation space should have smaller intra-class distance (higher tightness) and larger inter-class distance (higher looseness), we can measure the tightness and looseness levels of the learned representations with regard to ground-truth classes to verify whether the model has learned effective features for different classes from the unlabeled data. To this end, we propose two metrics named mean intra-Class Tightness (mACT) and mean inter-Class Looseness (mRCL) to quantitatively evaluate the merits of the learn skeleton representations.

Mean Intra-Class Tightness (mACT): To obtain mACT, we first compute intra-class tightness (ACT) of the \( k^{th} \) class with

\[
\text{ACT}(k) = \frac{D_{\text{global-average}}}{D_{k-\text{intra-class}}} = \frac{s_k}{C} \sum_{j=1}^{s_k} \sum_{i=1}^{C} \frac{1}{z} \sum_{j=1}^{s_k} D \left( v^k_j, c_z \right) \tag{12}
\]

where \( D_{k-\text{intra-class}} \) represents the average intra-class distance between the representation, \( v^k_j \), of \( j^{th} \) sample and the representation centroid, \( c_z \), belonging to the \( k^{th} \) class, \( D_{\text{global-average}} \) denotes the global average distance between each representation and all centroids, \( s_k \) represents the number of instances in the \( i^{th} \) class, \( C \) is the number of ground-truth classes, and \( D(v^k_j, c_z) = 1 - \frac{w_j \cdot c_z}{\| w_j \| \cdot \| c_z \|} \) is the cosine distance between \( v^k_j \) and \( c_z \). Note that \( 1/D_{k-\text{intra-class}} \) can be used to represent the \( k^{th} \)-class tightness, but this result does NOT consider the factor of global feature distribution that leads to different representation distances. Therefore, we use \( D_{\text{global-average}}/D_{k-\text{intra-class}} \) to indicate the degree of intra-class tightness. When \( D_{k-\text{intra-class}} \) is smaller than \( D_{\text{global-average}} \), the representations in \( k^{th} \) class are denser than the average, thus indicating a higher ACT of \( k^{th} \) class. Then, we average the ACT values of all classes to represent the mean intra-class tightness (mACT) of representations by

\[
mACT = \frac{1}{C} \sum_{k=1}^{C} \text{ACT}(k).
\]
Algorithm 1: Main algorithm of SPC-MGR

**Input:** Unlabeled training skeleton sequences $\Phi_t = \{S_{i,j}^t\}_{i=1}^{N_t}$, maximum distance $\epsilon$ and minimum sample amount $a_{\text{min}}$ for the DBSCAN algorithm, initialized embedding function $\psi(\cdot)$ for skeleton representation learning, temperature $\tau$

**Output:** Embedding function $\psi(\cdot)$

```plaintext
1 while not MaxEpoch do
   // Alternate clustering and contrastive learning
2   $\mathbf{M}_i = \psi(S_{i,j}^t)$ // Encode skeleton sequences to get instances $\{\mathbf{M}_i\}_{i=1}^{N_t}$
3   $\{\overline{M}^k\}_{k=1}^{M^o} = \text{DBSCAN}(\{\mathbf{M}_i\}_{i=1}^{N_t}, \epsilon, a_{\text{min}})$ // Generate clusters and discard outliers
4   $P^k = \text{Prototype}(\overline{M}^k)$ // Generate skeleton prototypes with Eq. 10
5   $\mathcal{L}_{\text{SPC}}(\{\mathbf{M}^k\}_{k=1}^{M^o}, \{P^k\}_{k=1}^{M^o}, \tau)$ // Calculate SPC loss with Eq. 11
6   update parameters of $\psi(\cdot)$ to minimize $\mathcal{L}_{\text{SPC}}$
7 end
```

Mean Inter-Class Looseness (mRCL): To measure the mRCL of representations, we compute the ratio of the average distance between different class centroids, $D_{\text{centroid-average}}$, to the global average distance, $D_{\text{global-average}}$ (see Eq. 12), with

$$\text{mRCL} = \frac{D_{\text{centroid-average}}}{D_{\text{global-average}}} = \frac{\sum_{i=1}^{C} \sum_{j=1}^{C} D(c_i, c_j)}{\sum_{i=1}^{C} \sum_{j=1}^{C} \sum_{z=1}^{C} D(v_j, c_z)}$$

(13)

where higher mRCL can be obtained with greater average distance $D_{\text{centroid-average}}$ between different representation centroids compared to the global average representation distance $D_{\text{global-average}}$. Full details of mACT and mRCL are provided in the Appendix. We conduct both quantitative and qualitative analysis of the proposed skeleton prototype contrastive learning scheme using the mACT and mRCL metrics, which shows that it can encourage the model to capture class-related semantics for more effective skeleton representations, as detailed in Sec. 5.4 and Appendix.

### 3.4 The Entire Approach

The computation flow of the proposed approach can be described as: $S \rightarrow G$ (Sec. 3.1) $\rightarrow F$ (Sec. 3.2) $\rightarrow M$ (Eq. 7) $\rightarrow \overline{M}$ (Sec. 3.3) $\rightarrow P$ (Eq. 10). For convenience, we use the embedding function $\psi(\cdot)$ to represent the multi-level skeleton graph representation encoding process, which can be formulated as $\psi(S) = \overline{M}$. We perform skeleton prototype contrastive learning by minimizing $\mathcal{L}_{\text{SPC}}$, so as to optimize $\psi(\cdot)$ and learn effective skeleton representations in an unsupervised manner, as illustrated in Algorithm 1. To facilitate better skeleton representation learning with more reliable clusters, we optimize our model by alternating clustering and contrastive learning. More technical details are provided in the Appendix. For the person re-ID task, we exploit the learned embedding function $\psi(\cdot)$ to encode each skeleton sequence of the probe set $\Phi_p$ into corresponding multi-level graph representation, $\{\overline{M}^{p,i}\}_{i=1}^{N_p}$, and match it with representations, $\{\overline{M}^{g,j}\}_{j=1}^{N_g}$, of the same identity in the gallery set $\Phi_g$ using Euclidean distance.

### 4 EXPERIMENTS

#### 4.1 Experimental Setup

We evaluate our approach on four skeleton-based person re-ID benchmarks: Kinect Gait Biometry Dataset (KGBD) [18], BIWI [26], KS20 VisLab Multi-View Kinect Skeleton Dataset [50], IAS-Lab [51], and a large-scale RGB video based multi-view gait dataset: CASIA-B [52]. They collect skeleton data from 164, 50, 20, 11, and 124 different individuals respectively, as detailed in Table 1. For BIWI and IAS-Lab datasets, we evaluate each testing set by setting it as the probe and the other one as the gallery. For KGBD, since no training and testing splits are given, we
randomly leave one skeleton video of each person as the probe set, and equally divide the remaining videos into the training set and gallery set. For KS20, we design different splitting setups to evaluate the multi-view person re-ID performance of our approach. In Random View Evaluation (RVE), we randomly select one sequence from each viewpoint as the probe sequence and use one half of the remaining skeleton sequences for training and the other half as the gallery. For Cross-View Evaluation (CVE), we divide probe samples by each viewpoint, including left lateral at 90°, left diagonal at 30°, frontal at 90°, right diagonal at 130°, and right lateral at 180°, to construct five individual viewpoint-based probe sets. We test each of them by matching identities from other four viewpoints to evaluate cross-view person re-ID performance. More details about the experimental setups are provided in the Appendix. Experiments with each evaluation setup are repeated for multiple times and the average performance is reported.

### 4.2 Implementation Details

The numbers of nodes in the part-level, body-level, and hyper-torsos-level graphs are \( n_1 = 10, n_2 = 5, \) and \( n_3 = 3 \) respectively. The sequence length \( f \) on four skeleton-based datasets (KS20, KGBD, BIWI, IAS-Lab) is empirically set to 6, which achieves best performance in average among different settings. For the largest dataset CASIA-B with roughly estimated skeleton data from RGB frames, we set sequence length \( f = 40 \) for training/testing. The node feature dimension is \( D_h = 8 \) and the number of structural relation heads is \( m = 16 \) for KGBD and \( m = 8 \) for other datasets. We use \( \lambda_a,b = 1 \) \((a, b \in \{1, 2, 3\})\) to averagely fuse multi-level graph features. For the DBSCAN algorithm, we empirically use maximum distance \( \epsilon = 0.6 \) (KGBD, BIWI), \( \epsilon = 0.8 \) (KS20, IAS-Lab), \( \epsilon = 0.75 \) (CASIA-B), and adopt minimum amount of samples \( a_{min} = 4 \) for KGBD and \( a_{min} = 2 \) for other datasets. We set the temperature \( \tau \) to 0.06 (KGBD), 0.075 (CASIA-B), 0.07 (BIWI), 0.08 (KS20, IAS-Lab) for skeleton prototype contrastive learning. We employ Adam optimizer with learning rate 0.00035 for all datasets. The batch size is 256 for KGBD and 128 for other datasets. For all methods compared in our experiments, we select optimal model parameters for training, and use their pre-defined skeleton descriptors or pre-trained skeleton representations for person re-ID. Full implementation details are provided in the Appendix and our codes are available at https://github.com/Kali-Hac/SPC-MGR.

### 4.3 Evaluation Metrics

We compute Cumulative Matching Characteristics (CMC) curve, which typically adopts top-1, top-5, and top-10 accuracy as the quantitative metrics, to show ratios that a probe sample matches the same identity in different-sized ranked lists of gallery samples. We also report Mean Average Precision (mAP) [53] to evaluate the overall performance of our approach. Unlike existing skeleton-based re-ID works that take only top-1 accuracy and nAUC measurements [54] under the assumption of fixed classes, this work evaluates all existing methods with more comprehensive metrics (top-1, top-5, top-10 accuracy and mAP) under the frequently-used person re-ID evaluation protocol, i.e., match IDs between probe and gallery, which can be extended to more general scenarios with varying classes such as generalized person re-ID tasks.

### 4.4 Performance Comparison

In this section, we compare our approach with state-of-the-art self-supervised and unsupervised skeleton-based person re-ID methods [19], [20], [22] on KS20, KGBD, IAS-Lab, and BIWI datasets with different probe settings in Table 2 and Table 3. As a reference for the overall performance, we include the latest supervised skeleton-based person re-ID methods [21], [25] and representative hand-crafted person re-ID methods [29], [30]. For deep learning based methods, we also report their model sizes, i.e., amount of network parameters, and computational complexity in Table 2.

#### 4.4.1 Comparison with Self-Supervised and Unsupervised Methods

As presented in Table 2 and Table 3, the proposed SPC-MGR enjoys distinct advantages over existing self-supervised and unsupervised methods on all datasets. Compared with AGE model [19] that learns skeleton features based on body-joint sequence representations, our approach consistently achieves higher person re-ID performance by a large margin of 17% over AGE model. Compared with SPC-MGR, our approach consistently achieves higher person re-ID performance by a large margin of 17% over AGE model.

#### 4.4.2 Comparison with Hand-Crafted Methods

Table 2: Performance comparison with existing hand-crafted, supervised, self-supervised, and unsupervised methods on KS20 (RVE setup), KGBD, and IAS-A testing sets. The amount of network parameters (million (M)) and computational complexity (giga floating-point operations (GFLOPs)) for the deep learning based methods are also reported. Bold numbers refer to the best cases among self-supervised/unsupervised methods.

| Types             | Methods               | # Params | GFLOPs | top-1 | top-5 | top-10 | mAP | top-1 | top-5 | top-10 | mAP | top-1 | top-5 | top-10 | mAP |
|-------------------|-----------------------|----------|--------|-------|-------|--------|-----|-------|-------|--------|-----|-------|-------|--------|-----|
| Hand-crafted      | \( D_{I3} \) Descriptors [26] | —        | —      | 39.4  | 71.7  | 81.7   | 18.9| 30.4  | 44.2  | 1.9   | 40.0 | 58.7  | 67.6  | 24.5   |
|                   | \( D_{I9} \) Descriptors [29] | —        | —      | 51.7  | 77.1  | 86.9   | 31.2| 50.9  | 59.8  | 4.0   | 42.7 | 62.9  | 70.7  | 25.2   |
| Supervised        | PoseGait [25]         | 8.95M    | 121.6M | 49.4  | 80.9  | 90.2   | 23.5| 50.6  | 67.0  | 7.2   | 55.7 | 79.0  | 86.2  | 21.7   |
|                   | MG-SCR [21]           | 0.35M    | 6.60M  | 46.3  | 75.4  | 84.0   | 10.4| 48.7  | 64.9  | 6.9   | 59.6 | 75.6  | 88.7  | 13.6   |
| Self-supervised   | SGEA [20]             | 8.47M    | 7.47M  | 45.0  | 65.0  | 75.1   | 38.1| 53.5  | 60.0  | 4.5   | 16.7 | 30.2  | 44.0  | 13.2   |
|                   | SM-SGE [22]           | 5.58M    | 22.61M | 45.9  | 71.9  | 81.1   | 9.5 | 54.2  | 60.7  | 4.4   | 36.4 | 59.6  | 69.5  | 14.1   |
|                   | SPC-MGR (Ours)        | 0.01M    | 0.12M  | 59.0  | 79.0  | 86.2   | 21.7| 48.7  | 64.9  | 6.9   | 59.6 | 75.6  | 88.7  | 13.6   |
| /Unsupervised     | Descriptors [29]      | —        | —      | —     | —     | —      | —   | —     | —     | —     | —    | —     | —     | —      |
|                   | — Descriptors [29]    | —        | —      | —     | —     | —      | —   | —     | —     | —     | —    | —     | —     | —      |

#### 4.4.3 Comparison with Existing Hand-Crafted, Supervised, and Self-Supervised Methods

Table 3: Performance comparison with existing hand-crafted, supervised, self-supervised, and unsupervised methods on KS20 (RVE setup), KGBD, and IAS-A testing sets. On two testing sets of BIWI, both SGEA and SM-SGE enjoy distinct advantages over existing self-supervised and unsupervised methods.
Finally, our approach also performs better than the latest graph-based skeleton representation learning method SM-SGE [22] by a distinct margin of 2.6-13.1% top-1 accuracy and 2.5-12.2% mAP on all datasets. In contrast to direct inter-sequence contrastive learning [20] or manually devising pretext tasks for skeleton representation learning [19], [22], our approach can automatically mine most representative skeleton features by contrasting sequence-level representations (instances) and cluster-level representations (prototypes), which enables our model to learn better skeleton representations for person re-ID. Moreover, our model requires only 0.01M parameters and evidently lower computational complexity for skeleton representation learning compared with existing self-supervised and unsupervised methods, as shown in Table 2, which demonstrates its superior efficiency for person re-ID tasks.

We also compare the performance of our approach with state-of-the-art skeleton-based counterparts with the cross-view evaluation (CVE) setup of KS20. As shown in Table 4, our approach remarkably outperforms the latest self-supervised and multi-view skeleton-based methods SGE [20] and SM-SGE [22] by an average margin of 6.5-32.1% top-1 accuracy, 12.8-41.0% top-5 accuracy, 17.5-40.0% top-10 accuracy, and 5.2-22.8% mAP on 24 out of 25 testing combinations of probe views and gallery views, which demonstrates that our model can learn more discriminative skeleton representations with better robustness against viewpoint variations for cross-view person re-ID.

### 4.4.2 Comparison with Hand-Crafted and Supervised Methods

Compared with $D_{13}$ [26] and $D_{16}$ [29] that extract hand-crafted geometric and anthropometric skeleton descriptors, our model achieves a significant improvement of person re-ID performance by 1.5-23.8% top-1 accuracy on KGBD, BIWI-S, and BIWI-W. Despite gaining similar performance on IAS-A and IAS-B, these methods are inferior to our approach by at least 7.3% top-1 accuracy on more challenging datasets such as KS20 and KGBD that contains more viewpoints and individuals. Furthermore, with unlabeled 3D skeletons as the only input, the proposed approach can obtain comparable or even superior performance to two state-of-the-art supervised methods PoseGait [25] and MG-SCR [21] on five out of six testing sets (KS20, IAS-A, IAS-B, BIWI-S, BIWI-W). Interestingly, with massive labels as the supervision, these methods still fail to obtain satisfactory person re-ID accuracy and even perform worse than hand-crafted methods on datasets with frequent view, shape, and appearance changes in KS20, IAS, and

### TABLE 3

| Types             | Methods                  | IAS-B | BIWI-S | BIWI-W |
|-------------------|--------------------------|-------|--------|--------|
|                   |                           | top-1 | top-5 | mAP    | top-1 | top-5 | mAP    | top-1 | top-5 | mAP    |
| Hand-crafted      | $D_{13}$ Descriptors [26] | 43.7  | 68.6  | 76.7  | 23.7  | 28.3  | 53.1  | 65.9  | 13.1  | 14.2  | 20.6  | 23.7  | 17.2  |
|                   | $D_{16}$ Descriptors [29] | 44.5  | 69.1  | 80.2  | 24.5  | 32.6  | 55.7  | 68.3  | 16.7  | 17.0  | 25.3  | 29.6  | 18.8  |
| Supervised        | PoseGait [25]             | 28.9  | 51.6  | 62.9  | 20.8  | 14.0  | 40.7  | 56.7  | 9.9   | 8.8   | 23.0  | 31.2  | 11.1  |
|                   | MG-SCR [21]               | 32.4  | 56.5  | 69.4  | 12.9  | 20.1  | 46.9  | 64.1  | 7.6   | 10.8  | 20.3  | 29.4  | 11.9  |
| Self-supervised   | AGE [19]                  | 31.1  | 52.3  | 64.2  | 12.8  | 25.1  | 43.1  | 61.6  | 8.9   | 11.7  | 21.4  | 27.3  | 12.6  |
| /Unsupervised     | SGEA [20]                 | 22.2  | 40.8  | 50.2  | 14.0  | 25.8  | 51.8  | 64.4  | 15.1  | 11.7  | 14.0  | 14.7  | 19.0  |
|                   | SM-SGE [22]               | 38.9  | 64.1  | 75.8  | 13.3  | 31.3  | 56.3  | 69.1  | 10.1  | 13.2  | 25.8  | 33.5  | 15.2  |
|                   | SPC-MGR (Ours)            | 43.3  | 68.4  | 79.4  | 24.1  | 34.1  | 57.3  | 69.8  | 16.0  | 18.9  | 31.5  | 40.5  | 19.4  |

### TABLE 4

| Gallery | Probe | 0° | 30° | 90° | 150° | 180° |
|---------|-------|----|-----|-----|------|------|
| Probe   |       |    |     |     |      |      |
| AGE [19]|       |    |     |     |      |      |
| SGEA [20]|      |    |     |     |      |      |
| SM-SGE  | [22]  |    |     |     |      |      |
| SPC-MGR (Ours)| |    |     |     |      |      |
| AGE     |       |    |     |     |      |      |
| SGEA    |       |    |     |     |      |      |
| SM-SGE  |       |    |     |     |      |      |
| SPC-MGR (Ours)| |    |     |     |      |      |

Frequently, view, shape, and appearance changes in KS20, IAS, and methods still fail to obtain satisfactory person re-ID accuracy and

views, which demonstrates that our model can learn more discriminative skeleton representations with better robustness against viewpoint variations for cross-view person re-ID.
Ablation study of our model with different components: Multi-level skeleton graphs (MG), multi-head structural relation layer (MSRL), full-level collaborative relation layer (FCRL), and skeleton prototype contrastive learning (SPC). “SG” denotes employing the single-level graph (part-level graph) and “+” indicates using the corresponding model component. “SG + MSRL” is evaluated under random model initialization without SPC.

### TABLE 5
| Id | Configurations       | KS20 Top-1 mAP | KGBD Top-1 mAP | IAS-A Top-1 mAP | IAS-B Top-1 mAP | BIWI-W Top-1 mAP | BIWI-S Top-1 mAP |
|----|----------------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 1  | Baseline             | 17.0           | 9.5            | 20.5           | 4.4            | 29.4           | 13.8           |
| 2  | SG + MSRL            | 18.6           | 10.2           | 21.4           | 3.7            | 30.5           | 14.3           |
| 3  | SG + MSRL + SPC      | 28.4           | 15.5           | 26.2           | 5.7            | 37.9           | 21.5           |
| 4  | MG + MSRL + SPC      | 45.1           | 21.2           | 34.5           | 6.3            | 40.0           | 22.5           |
| 5  | MG + MSRL + FCRL + SPC | 59.0         | 21.7           | 40.8           | 6.9            | 41.9           | 24.2           |

Performance of our approach on different datasets when exploiting different levels of graphs. “✓” indicates using corresponding graph level.

### TABLE 6
| Hyper-body-level | Body-level | Part-level | KS20 Top-1 mAP | KGBD Top-1 mAP | IAS-A Top-1 mAP | IAS-B Top-1 mAP | BIWI-W Top-1 mAP | BIWI-S Top-1 mAP |
|------------------|------------|------------|----------------|----------------|----------------|----------------|----------------|----------------|
| ✓                | ✓          | ✓          | 29.9           | 15.6           | 15.2           | 4.2            | 30.6           | 20.4           |
| ✓                | ✓          | ✓          | 46.1           | 20.5           | 26.7           | 5.7            | 39.3           | 21.3           |
| ✓                | ✓          | ✓          | 53.3           | 21.3           | 35.6           | 6.3            | 41.7           | 21.9           |
| ✓                | ✓          | ✓          | 59.0           | 21.7           | 40.8           | 6.9            | 41.9           | 24.2           |

| m     | KS20 Top-1 mAP | KGBD Top-1 mAP | IAS-A Top-1 mAP | IAS-B Top-1 mAP | BIWI-W Top-1 mAP | BIWI-S Top-1 mAP |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|
| 2     | 52.7           | 20.1           | 30.3           | 6.2            | 38.9           | 22.2           |
| 4     | 53.9           | 21.4           | 33.7           | 6.5            | 39.3           | 19.4           |
| 8     | 59.0           | 21.7           | 38.9           | 6.6            | 41.9           | 24.2           |
| 16    | 56.5           | 21.6           | 40.8           | 6.9            | 41.0           | 23.4           |

| m     | KS20 Top-1 mAP | KGBD Top-1 mAP | IAS-A Top-1 mAP | IAS-B Top-1 mAP | BIWI-W Top-1 mAP | BIWI-S Top-1 mAP |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|
| 2     | 52.7           | 20.1           | 30.3           | 6.2            | 38.9           | 22.2           |
| 4     | 53.9           | 21.4           | 33.7           | 6.5            | 39.3           | 19.4           |
| 8     | 59.0           | 21.7           | 38.9           | 6.6            | 41.9           | 24.2           |
| 16    | 56.5           | 21.6           | 40.8           | 6.9            | 41.0           | 23.4           |

Performance of our approach on different datasets when employing different amounts of structural relation heads (m = 2, 4, 8, 16).

### TABLE 7
| m     | KS20 Top-1 mAP | KGBD Top-1 mAP | IAS-A Top-1 mAP | IAS-B Top-1 mAP | BIWI-W Top-1 mAP | BIWI-S Top-1 mAP |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|
| 2     | 68.1           | 26.8           | 53.9           | 12.1           | 54.2           | 23.7           |
| 4     | 70.0           | 27.6           | 55.3           | 12.8           | 55.5           | 24.4           |
| 8     | 71.5           | 28.4           | 57.0           | 13.4           | 56.5           | 25.1           |
| 16    | 7,55.0         | 29.2           | 58.5           | 14.0           | 57.5           | 25.8           |

Performance of our approach on different datasets when setting different coefficients (λC = 0.1, 0.25, 0.5, 1.0) for multi-level graph feature fusion.

### TABLE 8
| λC   | KS20 Top-1 mAP | KGBD Top-1 mAP | IAS-A Top-1 mAP | IAS-B Top-1 mAP | BIWI-W Top-1 mAP | BIWI-S Top-1 mAP |
|------|----------------|----------------|----------------|----------------|----------------|----------------|
| 1.0  | 68.1           | 26.8           | 53.9           | 12.1           | 54.2           | 23.7           |
| 0.25 | 63.9           | 25.5           | 53.5           | 11.8           | 53.7           | 23.4           |
| 0.5  | 55.5           | 24.7           | 49.9           | 11.1           | 49.0           | 22.8           |
| 1.0  | 50.0           | 24.0           | 47.6           | 10.6           | 45.7           | 21.7           |

BIWI with the probe-gallery person re-ID setup. Considering that our approach does not require any manual annotation and can achieve highly competitive and more balanced performance with a significantly smaller size of network parameters, as shown in Table 2, it can be a more general solution to skeleton-based person re-ID and related tasks. We will show broader applications of our approach to more general person re-ID scenarios in Sec. 5.5.

## 5 Further Analysis

### 5.1 Ablation Study

In this section, we conduct ablation study to demonstrate the necessity of each component in the proposed approach. We use raw skeleton sequences as the baseline and report average performance of each configuration. As reported in Table 5, we can draw the following conclusions. The model utilizing single-level skeleton graph with MSRL (Id = 2, 3) shows higher performance than the baseline (Id = 1) that directly uses raw body-joint sequences by 0.3-8.5% top-1 accuracy and 0.7-7.7% mAP on all datasets, regardless of using SPC. Such results demonstrate the effectiveness of graph representations, as it can model richer body structural information and mine valuable body-component relations to obtain a more discriminative skeleton representation. Compared with the model without contrastive learning (Id = 2), employing SPC (Id = 3) obtains consistent re-ID performance improvement by up to 9.8% top-1 accuracy and 7.5% mAP on different datasets. This justifies that the proposed SPC is a highly effective contrastive learning paradigm, which enables the model to mine more typical and unique skeleton features of different identities from the unlabeled graph representations for person re-ID. Exploiting the proposed multi-level graphs (Id = 4) performs better than solely using single-level graph (Id = 3) with a remarkable margin of 2.1-16.7% top-1 accuracy and 0.4-5.7% mAP, which demonstrates that modeling body structure and relations at various levels with the proposed graph representations (MG) can encourage the model to learn more useful skeleton...
features for person re-ID. We will compare the effects of different level graphs in the next section as well. Adding FCRL further improves the overall performance in terms of both top-1 accuracy by 0.8-13.9% and mAP by 0.5-2.6% on different datasets. Such results verify our claim that combining structural and collaborative body relation learning can facilitate capturing richer features of body structure and skeleton patterns for the person re-ID task.

5.2 Discussions

5.2.1 Different Skeleton Graph Levels

To analyze the effects of different graph levels, Table 6 evaluates the performance of our approach when combining different graphs for person re-ID under the same optimal model setting. Exploiting low-level graph representations can achieve evidently better person re-ID performance than using high-level graphs by 1.3-19.0% top-1 accuracy on most datasets, which suggests that the low-level graphs (e.g., part-level) that contain more body-component nodes and body structural/relational information can benefit learning more discriminative skeleton representations for person re-ID. Combining different levels of graphs can improve the person re-ID performance compared with solely using single-level graphs (first three rows in Table 6) in most cases, while the model exploiting all level graphs is the best performer among different datasets. This further justifies that the proposed multi-level graphs that model body structure and motion at various levels enable our approach to capture more recognizable structural features and pattern information for the person re-ID task.

5.2.2 Different Structural Relation Head Settings

As shown in Table 7, we show the effects of different amounts of structural relation heads on our approach. It is observed that introducing more learnable structural relation heads with larger m values can improve the performance of our approach on different datasets, as it can encourage capturing more spatially or motion related features of adjacent body components. However, too many structural relation heads, m = 16, for example, may cause the model to learn redundant relation information and degrade model performance on relatively limited training data such as KS20.

5.2.3 Different Graph Fusion Coefficients

The hyper-parameter λ controls the extent of fusing collaborative node features between \( g^h \)th level and \( l^h \)th level graphs. Here we unify \( \lambda \) to equally fuse features between different graph levels, so as to better analyze the overall effects of collaborative fusion. As reported in Table 8, improving \( \lambda \) enables our model to achieve better Re-ID performance in terms of both top-1 accuracy and mAP on all datasets. Such results verify the necessity of adequate collaborative graph fusion to learn more effective multi-level skeleton graph representations for person re-ID. Based on this observation, we set \( \lambda_{a,b}^{1,2} = 1 \) to fuse skeleton graph features.

5.2.4 Temperature Settings for SPC Scheme

As shown in Table 9, the model using lower temperature \( \tau \) can obtain slightly better person re-ID performance. Since higher temperatures tend to retain more similar features between different skeleton prototypes, it might reduce the capacity of contrastive learning on capturing inter-prototype differences and their underlying discriminative features. This result is also consistent with our previous analysis that the proposed SPC scheme under an appropriate temperature can encourage learning higher mRCL, i.e., inter-class differences, and more effective skeleton representations for the person re-ID task. We also found that a high temperature

| \( \tau \) | KS20 mAP | KGBD mAP | IAS-A mAP | IAS-B mAP | BIWI-W mAP | BIWI-S mAP |
|---|---|---|---|---|---|---|
| 0.06 | 58.3 | 21.2 | 40.8 | 6.9 | 41.2 | 22.5 | 43.2 | 24.0 | 18.5 | 18.7 | 34.5 | 15.6 |
| 0.07 | 58.6 | 21.3 | 38.6 | 6.6 | 41.5 | 22.7 | 43.0 | 23.8 | 18.9 | 19.4 | 34.1 | 16.0 |
| 0.08 | 59.0 | 21.7 | 36.6 | 6.6 | 41.9 | 24.2 | 43.3 | 24.1 | 18.8 | 19.9 | 33.9 | 14.5 |
| 0.1 | 57.7 | 21.6 | 34.0 | 6.5 | 41.4 | 23.1 | 43.8 | 24.2 | 18.6 | 17.6 | 32.8 | 12.9 |
| 0.25 | 57.0 | 21.5 | 27.7 | 5.7 | 41.0 | 21.4 | 43.2 | 23.5 | 18.3 | 16.2 | 33.8 | 15.1 |
| 0.5 | 56.5 | 21.1 | 19.8 | 5.0 | 40.8 | 24.0 | 41.0 | 24.5 | 19.0 | 18.5 | 33.0 | 15.0 |

| \( \alpha_{min} \) | KS20 mAP | KGBD mAP | IAS-A mAP | IAS-B mAP | BIWI-W mAP | BIWI-S mAP |
|---|---|---|---|---|---|---|
| 1 | 53.9 | 21.5 | 30.3 | 6.3 | 41.2 | 23.1 | 43.0 | 23.6 | 18.4 | 18.5 | 34.0 | 14.3 |
| 2 | 59.0 | 21.7 | 34.4 | 6.5 | 41.9 | 24.2 | 43.3 | 24.1 | 18.9 | 19.4 | 34.1 | 16.0 |
| 3 | 58.4 | 21.2 | 36.6 | 6.5 | 41.1 | 22.6 | 43.2 | 23.9 | 19.1 | 19.3 | 32.0 | 13.8 |
| 4 | 58.6 | 21.5 | 40.8 | 6.9 | 40.0 | 22.1 | 42.8 | 24.2 | 18.8 | 18.1 | 30.9 | 13.3 |

| \( \epsilon \) | KS20 mAP | KGBD mAP | IAS-A mAP | IAS-B mAP | BIWI-W mAP | BIWI-S mAP |
|---|---|---|---|---|---|---|
| 0.4 | 53.3 | 21.7 | 33.9 | 5.8 | 39.8 | 23.1 | 43.2 | 24.3 | 18.6 | 17.6 | 24.6 | 11.2 |
| 0.6 | 55.7 | 21.5 | 40.8 | 6.9 | 41.3 | 22.6 | 42.7 | 24.0 | 18.9 | 19.4 | 34.1 | 16.0 |
| 0.8 | 59.0 | 21.7 | 22.1 | 5.4 | 41.9 | 24.2 | 43.3 | 24.1 | 17.3 | 18.5 | 33.2 | 14.4 |
| 1.0 | 54.3 | 21.0 | 17.7 | 4.0 | 40.7 | 22.2 | 41.0 | 23.8 | 12.3 | 15.1 | 28.9 | 12.0 |
could lead to an evident degradation of model performance on the KGBD dataset that contains more individuals. In practice, we select the best model temperature for datasets of different sizes to achieve more balanced and better person re-ID performance.

5.2.5 Different Settings for DBSCAN
We adjust the sample density, i.e. minimum sample amount \( a_{\text{min}} \), with the maximum distance \( \epsilon \), in the DBSCAN algorithm to encourage more balanced and stable clustering. In Table 10 and Table 11, it is seen that relatively lower \( a_{\text{min}} \) and higher \( \epsilon \), which enhances the connectedness of skeleton instances in larger clusters and generates less prototypes, can boost the model performance in most cases. However, on larger datasets such as KGBD, setting a high \( \epsilon \) is shown to reduce the overall performance, while adopting lower \( \epsilon \) and higher \( a_{\text{min}} \), i.e., improving the amount of prototypes, can facilitate person re-ID performance. It suggests that more diverse skeleton prototypes may encourage mining richer discriminative features in the case of more identities.

5.3 Analysis of Body Relation Learning
To validate the ability of our approach on inferring body-component relations between different graphs, we visualize collaborative relations and corresponding relation value matrices of different graphs in Fig. 6. Here we take collaborative relations between adjacent graph levels as an example for analysis. As shown in Fig. 6 (a)-(c), we can see that our approach can learn to assign larger relation values to significantly moving body portions such as feet, hips, hands, and elbows (nodes 0, 1, 2, 3, 6, 7, 8 and 9 in \( G_1 \)), which typically perform collaborative motion during walking process and can carry abundant and unique motion features. By inferring inherent relations between these body components and fusing them into multi-level graph representations, our model is able to focus on crucial patterns of skeletal motion to learn more effective skeleton representations. Our approach can not only capture relations of major body components to their corresponding high-level components, but also allocate different attention to potentially motion-related components. For example, as shown in Fig. 6(d), Fig. 6(f) and Fig. 6(h), the movement of arms (nodes 3 and 4 in \( G_2 \)) is highly correlated with the leg motion (nodes 0, 1, 2 and 3 in \( G_1 \)), and different parts of them enjoy varying degree of collaboration in walking. Such results demonstrate that our approach can capture body-related high-level semantics (e.g., body-component correspondence) and meanwhile facilitate global pattern learning of correlative components. More graph details and visualization results are provided in the Appendix.

5.4 Analysis of Skeleton Representations
5.4.1 Intra-Class Tightness and Inter-Class Looseness
In this section, we further analyze the effectiveness of the learned skeleton representations with the proposed mACT and mRCL metrics, and compare our approach with existing self-supervised and unsupervised methods, AGE [19], SGELA [20], SM-SGE [22], on all datasets. As shown in Fig. 7, we can obtain the following crucial analysis. The representations learned by the proposed SPC-MGR enjoy the highest mACT and mRCL among all methods on different datasets, which indicates that our approach is able to learn higher similarity for the same-class instances while improving the distance between different classes. Such results also substantiate our intuition on SPC scheme that maximizing intra-prototype similarity and inter-prototype dissimilarity can encourage capturing both similar and unique semantics of ground-truth classes in skeleton representation learning. The model obtaining both higher mACT and mRCL can also achieve better person re-ID performance in most cases. For example, our method, SGELA, and SM-SGE achieve higher top-1 accuracy and mAP than AGE on KS20 and BIWI testing sets (see Table 2 and Table 3), which is consistent with results of mACT and mRCL shown in Fig. 7(b) and Fig. 7(d). This further verifies that the proposed mACT and mRCL can serve as auxiliary evaluation metrics to help analyze the effectiveness of learned representations in person re-ID tasks.

5.4.2 Visualization of Skeleton Representations
We conduct a T-SNE [59] visualization of skeleton representations for a qualitative analysis, and compare our approach with two state-of-the-art skeleton-based methods, i.e., AGE [19], SM-SGE [22]. As presented in Fig. 8(c), the skeleton representations learned...
Testing mACT

Testing mRCL

SM-SGE (BIWI)

SPC-MGR (KS20)

AGE (KS20)

SM-SGE (KS20)

Fig. 7. Mean intra-class tightness (mACT) and mean inter-class looseness (mRCL) comparison with existing self-supervised and unsupervised methods (AGE, SGEA, SM-SGE) on different datasets. Note: We report the average mACT and mRCL of all testing sets of BIWI and IAS-Lab.

Testing mACT

Testing mRCL

Fig. 8. T-SNE visualization of the skeleton representations learned by AGE (a), (d)), SM-SGE (b), (e)), and the proposed SPC-MGR ((c), (f)) for the first 10 classes in BIWI and KS20 datasets. Note: Different colors indicate skeleton representations of different classes.

Table 12

Performance comparison with appearance-based and skeleton-based methods on CASIA-B. Note: “Cl-Nm” denotes the probe set under “Clothes” condition and gallery set under “Normal” condition. † refers to appearance-based methods and * represents requiring label information for training. “—” indicates no published result.

| Method                          | Probe-Gallery | Nm-Nm | Nm-Bg | Bg-Bg | Cl-Cl | Cl-Nm | Bg-Nm |
|--------------------------------|---------------|-------|-------|-------|-------|-------|-------|
|                                | top-1         | top-5 | top-10| mAP   | top-1 | top-5 | top-10| mAP   | top-1 | top-5 | top-10| mAP   |
| 1LMNN† [55]                    | 3.9           | 22.7  | 36.1  | 18.3  | 38.6  | 49.2  | 17.4  | 35.7  | 45.8  | 11.6  | 12.6  | 17.8  | 23.1  | 37.1  | 44.4  |
| 1ITML† [56]                    | 7.5           | 22.2  | 34.2  | 19.5  | 26.0  | 33.7  | 20.1  | 34.4  | 43.3  | 10.3  | 24.5  | 36.1  | 21.8  | 30.4  | 36.3  |
| 1ELF† [54]                     | 12.3          | 35.6  | 50.3  | 5.8   | 25.5  | 37.6  | 19.9  | 43.9  | 56.7  | 5.6   | 16.0  | 26.3  | 17.1  | 30.0  | 37.9  |
| 1DALF† [57]                    | 4.9           | 27.0  | 41.6  | 10.2  | 33.5  | 47.2  | 16.7  | 42.0  | 56.7  | 11.6  | 19.4  | 27.6  | 22.9  | 30.1  | 36.1  |
| 1Score-based MLR† [58]         | 13.6          | 48.7  | 63.7  | 13.6  | 48.7  | 63.7  | 13.5  | 48.6  | 63.9  | 9.7   | 27.8  | 45.1  | 14.7  | 32.6  | 50.2  |
| 1Feature-based MLR† [58]       | 16.3          | 43.4  | 60.8  | 18.9  | 44.8  | 59.4  | 25.4  | 53.3  | 68.9  | 20.3  | 42.6  | 56.9  | 31.8  | 53.6  | 64.1  |
| AGE [19]                       | 20.8          | 29.3  | 34.2  | 3.5   | 37.1  | 56.2  | 67.0  | 9.8   | 35.5  | 54.3  | 65.3  | 9.6   | 14.6  | 33.0  | 42.7  | 3.0   |
| SM-SGE [22]                    | 50.2          | 73.5  | 81.9  | 6.6   | 26.6  | 49.0  | 59.4  | 9.3   | 27.2  | 51.4  | 63.2  | 9.7   | 10.6  | 26.3  | 35.9  | 3.0   |
| SPC-MGR (Ours)                 | 71.2          | 88.0  | 92.8  | 9.1   | 44.3  | 66.4  | 76.4  | 11.4  | 48.5  | 71.6  | 81.8  | 11.8  | 22.4  | 40.4  | 51.0  | 4.3   |

by our approach can form different class clusters with higher separation than AGE and SM-SGE on BIWI, which suggests the lower entropy of our representations. Interestingly, it is observed that the learned representations on KS20 are separated in small groups of the same class, as shown in Fig. 8(f), which enjoys significantly larger looseness than other two methods. Such results imply that our model may learn skeleton representations with finer separation and enable pattern-based grouping in a specific class.

5.5 Broader Applications Under General Scenarios

In more general person re-ID scenarios such as large-scale RGB-based scenes, the source datasets might only contain RGB images without any skeleton data or lack sufficient data for model training. For the former case, the proposed approach can exploit unlabeled 3D skeletons estimated from RGB videos to learn effective representations for the person re-ID task. In the later case, our approach is able to directly transfer the pre-trained model to perform person re-ID on a new dataset. We show the broader applications of our approach under these general settings as follows.

5.5.1 Application to Model-Estimated Skeleton Data

To verify the effectiveness of our skeleton-based approach when applied to large-scale RGB-based settings (CASIA-B), we exploit pre-trained pose estimation models [60], [61] to extract 3D skeleton data from RGB videos of CASIA-B, and evaluate the performance of our approach with the estimated skeleton data. We compare our approach with representative appearance-based methods [54]–[58] and skeleton-based methods [19], [22]. As reported in Table 12, our approach is superior to recent skeleton-based methods SM-SGE and AGE with an evident performance gain of 7.2–50.4% top-1 accuracy and 1.1–5.6% mAP in four out of five evaluation conditions of CASIA-B, which substantiates that the proposed approach is capable of learning more discriminative skeleton representations than these methods in the case of using model-estimated skeleton data. Compared with representative classic appearance-based methods that utilize visual features, e.g., RGB features and silhouettes, our skeleton-based approach still achieves the best performance in most conditions. For instance, our approach not only performs better than LMNN [55] and ITML
Generalized person re-ID performance of our approach with direct domain generalization (DG) from source datasets (“Source”) to target datasets (“Target”). “UF” represents fine-tuning the source model with the unlabeled data of target datasets. BIWI-W/S denotes the Walking/Still testing set of BIWI. Bold numbers indicates that the model using “DG” or “UF” obtains better performance than the original one trained on the same dataset.

| Source | KS20 | KGBD | IAS-Lab | BIWI |
|--------|------|------|---------|------|
| Target | Type | top-1 | top-5 | top-10 | mAP | top-1 | top-5 | top-10 | mAP | top-1 | top-5 | top-10 | mAP |
| KS20   | DG   | 19.7 | 54.3 | 69.7 | 10.2 | 20.1 | 60.4 | 75.6 | 13.5 | 29.7 | 62.7 | 79.9 | 15.0 |
|        | UF   | 59.0 | 79.0 | 86.2 | 21.7 | 48.4 | 77.7 | 85.2 | 21.6 | 52.2 | 78.3 | 89.1 | 22.8 |
| KGBD   | DG   | 18.3 | 37.3 | 46.7 | 4.4  | 21.2 | 39.6 | 50.8 | 17.0 | 22.0 | 41.2 | 51.9 | 13.0 |
|        | UF   | 28.5 | 45.2 | 52.9 | 6.4  | 40.8 | 57.5 | 65.0 | 6.9  | 31.1 | 48.1 | 55.7 | 6.5  |
| IAS-A  | DG   | 27.9 | 57.1 | 71.5 | 15.6 | 29.6 | 56.5 | 71.6 | 16.0 | —   | —   | —   | 27.5 |
|        | UF   | 42.8 | 67.7 | 77.5 | 23.0 | 34.1 | 59.4 | 72.0 | 18.4 | 41.9 | 66.3 | 75.6 | 24.2 |
| IAS-B  | DG   | 32.0 | 60.6 | 72.0 | 15.7 | 29.5 | 58.8 | 68.9 | 16.5 | —   | —   | —   | 27.0 |
|        | UF   | 45.4 | 68.8 | 81.4 | 29.0 | 35.9 | 61.6 | 72.1 | 21.9 | 43.3 | 68.4 | 79.4 | 24.1 |
| BIWI-W | DG   | 19.2 | 31.5 | 38.9 | 19.6 | 10.3 | 22.5 | 33.1 | 12.0 | 10.0 | 23.1 | 31.1 | 15.9 |
|        | UF   | 21.6 | 32.3 | 40.9 | 20.7 | 15.0 | 29.6 | 39.1 | 14.3 | 17.7 | 29.3 | 36.2 | 16.6 |
| BIWI-S | DG   | 23.8 | 52.2 | 69.9 | 14.2 | 19.0 | 49.4 | 67.4 | 8.5  | 18.8 | 46.1 | 61.1 | 12.2 |
|        | UF   | 40.4 | 62.9 | 74.2 | 16.2 | 21.3 | 46.5 | 57.2 | 9.8  | 27.9 | 44.5 | 63.9 | 12.8 |

6 Conclusion and Future Work

In this paper, we devise unified multi-level graphs to represent 3D skeletons, and propose an unsupervised skeleton prototype contrastive learning paradigm with multi-level relation modeling (SPC-MGR) to learn effective skeleton representations for person re-ID. We devise a multi-head structural relation layer to capture relations of neighbor body-component nodes in graphs, so as to aggregate key correlative features into effective node representations. To capture more discriminative patterns in skeletal motion, we propose a full-level collaborative relation layer to infer dynamic collaboration among different-level components. Meanwhile, a multi-level graph fusion is exploited to integrate collaborative node features across graphs to enhance structural semantics and global pattern learning. Lastly, we propose a skeleton prototype contrastive learning scheme to cluster unlabeled skeleton graph representations and contrast their inherent similarity with representative skeleton features to learn effective skeleton representations for person re-ID. The proposed SPC-MGR outperforms several state-of-the-art skeleton-based methods, and is also highly effective in more general person re-ID scenarios.

There exist two limitations in our study. The 3D skeletons used in this work are collected with relatively high precision with little noise, while a more general scenario for skeleton collection and person re-ID (e.g., extracted from RGB videos in outdoor scenes) has not been thoroughly studied. The datasets for evaluation are with relatively limited scale when compared with prevalent RGB-based re-ID datasets like MSMT17, since large-scale 3D skeleton based re-ID benchmarks with more pedestrians and scenarios are still unavailable. To facilitate skeleton-based research, we will build and open our own skeleton-based re-ID datasets in the future.

We believe that this work makes progress towards lightweight, general, and robust person re-ID research, where we for the first time exploit unlabeled 3D skeleton data of small scale to learn discriminative and generalizable pedestrian representations to effectively perform unsupervised, multi-view, and generalized person re-ID. There are several potential directions to be further explored. More efficient clustering schemes, such as memory-based clustering, could be devised to improve the stability and consistency of skeleton representation learning in clustering. Combining finer-grained contrastive learning is able to help bootstrap clustering, while employing skeleton augmentation strategies could sample...
more instances to enhance the skeleton prototype contrastive learning. Another important direction is to explore graph-based self-supervised pretext tasks to facilitate capturing richer high-level graph semantics from unlabeled graph representations. Our model can be potentially transferred to other skeleton-based tasks, such as 3D human action recognition, and we can expect it to synergize other modalities for more computer vision tasks.

7 Ethical Statements
Person re-ID is a crucial task providing significant value for both academia and industry. Nevertheless, it could be a controversial technology like face recognition, since its improper application will probably pose threat to the public privacy and society security. In this context, we would like to emphasize that all datasets in our experiments are either publicly available (IAS-Lab, BIWI, KGBD) or officially authorized (CASIA-B, KS20). The official agents of those datasets have confirmed and guaranteed that all data are collected, processed, released, and used with the consent of subjects. For the protection of privacy, all individuals in datasets are anonymized with simple identity numbers. Besides, our approach and models must only be used for research purpose.

References
[1] A. Nambiar, A. Bernardino, and J. C. Nascimento, “3D-based person re-identification: A survey,” ACM Computing Surveys, vol. 52, no. 2, p. 33, 2019.
[2] W.-S. Zheng, S. Gong, and T. Xiang, “Towards open-world person re-identification by one-shot group-based verification,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 3, pp. 591–606, 2015.
[3] D. Baltieri, R. Vezzani, and R. Cucchiara, “Sarc3d: a new 3d body model for people tracking and re-identification,” in International Conference on Image Analysis and Processing. Springer, 2011, pp. 197–206.
[4] R. Vezzani, D. Baltieri, and R. Cucchiara, “People reidentification in surveillance and forensics: A survey,” ACM Computing Surveys, vol. 46, no. 2, p. 29, 2013.
[5] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, “Multi-task learning with low rank attribute embedding for multi-camera person re-identification,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 5, pp. 1167–1181, 2018.
[6] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, “Person re-identification by camera correlation aware feature augmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 2, pp. 392–408, 2018.
[7] M. Li, X. Zhu, and S. Gong, “Unsupervised tracklet person re-identification,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–12, 2019.
[8] X. Qian, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue, “Leader-based multi-scale attention deep architecture for person re-identification,” IEEE Transaction on Pattern Analysis and Machine Intelligence, 2019.
[9] H.-X. Yu, A. Wu, and W.-S. Zheng, “Unsupervised person re-identification by deep asymmetric metric embedding,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 956–973, 2020.
[10] L. Wang, T. Tan, H. Ning, and W. Hu, “Silhouette analysis-based gait recognition for human identification,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 12, pp. 1505–1518, 2003.
[11] C. Wang, J. Zhang, L. Wang, J. Fu, and X. Yuan, “Human identification using temporal information preserving gait template,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pp. 2164–2176, 2011.
[12] T. Wang, S. Gong, X. Zhu, and S. Wang, “Person re-identification by discriminative selection in video ranking,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 12, pp. 2501–2514, 2016.
[13] R. Zhao, W. Oyang, and X. Wang, “Person re-identification by saliency learning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 2, pp. 356–370, 2017.
[14] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, “Densely semantically aligned person re-identification,” in CVPR, 2019, pp. 667–676.
[15] N. Karianakis, Z. Liu, Y. Chen, and S. Soatto, “Reinforced temporal attention and split-rate transfer for depth-based person re-identification,” in ECCV, 2018, pp. 715–733.
[16] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, “Self-paced contrastive learning with hybrid memory for domain adaptive object re-id,” in NeurIPS, 2020.
[17] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, “Re-identification with rgb-d sensors,” in ECCV. Springer, 2012, pp. 433–442.
[18] V. O. Andersson and R. M. Araujo, “Person identification using anthropometric and gait data from Kinect sensor,” in AAAI, 2015.
[19] H. Rao, S. Wang, X. Hu, M. Tan, H. Da, J. Cheng, and B. Hu, “Self-supervised gait encoder with locality-aware attention for person re-identification,” in IJCAI, vol. 1, 2020, pp. 898–905.
[20] 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, 2021.
[21] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, “Multi-level graph encoding with structural-collaborative relation learning for skeleton-based person re-identification,” in IJCAI, 2021.
[22] H. Rao, X. Hu, J. Cheng, and B. Hu, “Sm-sge: A self-supervised multi-scale skeleton graph encoding framework for person re-identification,” in Proceedings of the 29th ACM international conference on Multimedia, 2021.
[23] F. Han, B. Reily, W. Hoff, and H. Zhang, “Space-time representation of people based on 3d skeleton data: A review,” Computer Vision and Image Understanding, vol. 158, pp. 85–105, 2017.
[24] R. Tanawongsuwan and A. Bobick, “Gait recognition from time-normalized joint-angle trajectories in the walking plane,” in CVPR, vol. 2, Dec 2001, pp. II–II.
[25] R. Liao, S. Yu, W. An, and Y. Huang, “A model-based gait recognition method with body pose and human prior knowledge,” Pattern Recognition, vol. 98, p. 107069, 2020.
[26] M. Munaro, A. Fossati, A. Basso, E. Menegatti, and L. Van Gool, “One-shot person re-identification with a consumer depth camera,” in Person Re-Identification. Springer, 2014, pp. 161–181.
[27] J.-H. Yoo, M. S. Nixon, and C. J. Harris, “Extracting gait signatures based on anatomical knowledge,” in Proceedings of BMVA Symposium on Advancing Biometric Technologies. Citeseer, 2002, pp. 596–606.
[28] M. P. Murray, A. B. Drought, and R. C. Kory, “Walking patterns of normal men,” Journal of Bone and Joint Surgery, vol. 46, no. 2, pp. 335–360, 1964.
[29] P. Pala, L. Seidenari, S. Berretti, and A. Del Bimbo, “Enhanced skeleton and face 3d data for person re-identification from depth cameras,” Computers & Graphics, 2019.
[30] M. Munaro, A. Basso, A. Fossati, L. Van Gool, and E. Menegatti, “3d reconstruction of freely moving persons for re-identification with a depth sensor,” in IJCAI, 2014, pp. 4512–4519.
[31] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in CVPR, vol. 2, 2006, pp. 1735–1742.
[32] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in CVPR, 2018, pp. 3733–3742.
[33] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” arXiv preprint arXiv:1807.03748, 2018.
[34] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, “Big unsupervised models are strong semi-supervised learners,” in NeurIPS, 2020.
[35] J. Li, P. Zhou, C. Xiong, and S. Hoi, “Prototypical contrastive learning of unsupervised representations,” in ICLR, 2021.
[36] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, “Discriminative unsupervised feature learning with convolutional neural networks,” in NeurIPS, 2014, pp. 766–774.
[37] M. Gutmann and A. Hyvärinen, “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models,” in International Conference on Artificial Intelligence and Statistics, 2010, pp. 297–304.
[38] C. Zhuang, A. L. Zhai, and D. Yamins, “Local adaptation of unsupervised learning of visual embeddings,” in ICCV, 2019, pp. 6002–6012.
[39] Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” arXiv preprint arXiv:1906.05849, 2019.
[40] I. Misra and L. van der Maaten, “Self-supervised learning of pretext-invariant representations,” in CVPR, 2020, pp. 6707–6717.
[41] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, “Unsupervised embedding learning via invariant and spreading instance feature,” in CVPR, 2019, pp. 6210–6219.
[42] X. Ji, J. F. Henriques, and A. Vedaldi, “Invariant information clustering for unsupervised image classification and segmentation,” in ICCV, 2019, pp. 9865–9874.

[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in ICML, 2020.

[44] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in CVPR, 2020, pp. 9729–9738.

[45] D. A. Winter, Biomechanics and motor control of human movement. John Wiley & Sons, 2009.

[46] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata, “Nonrigid motion analysis: Articulated and elastic motion,” Computer Vision and Image Understanding, vol. 70, no. 2, pp. 142–156, 1998.

[47] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks.” in ICLR, 2018.

[48] 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 KDD, vol. 96, no. 34, 1996, pp. 226–231.

[49] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in NeurIPS Workshop, 2014.

[50] A. Nambiar, A. Bernardino, J. C. Nascimento, and A. Fred, “Context-aware person re-identification in the wild via fusion of gait and anthropometric features,” in International Conference on Automatic Face & Gesture Recognition. IEEE, 2017, pp. 973–980.

[51] M. Munaro, S. Ghidoni, D. T. Dizmen, and E. Menegatti, “A feature-based approach to people re-identification using skeleton keypoints,” in ICRA. IEEE, 2014, pp. 5644–5651.

[52] S. Yu, D. Tan, and T. Tan, “A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition,” in ICPR, 2006, pp. 441–444.

[53] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in ICCV, 2015, pp. 1116–1124.

[54] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in ECCV. Springer, 2008, pp. 262–275.

[55] K. Q. Weinberger and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification.” Journal of machine learning research, vol. 10, no. 2, 2009.

[56] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning.” in ICML, 2007, pp. 209–216.

[57] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in CVPR. IEEE, 2010, pp. 2360–2367.

[58] Z. Liu, Z. Zhang, Q. Wu, and Y. Wang, “Enhancing person re-identification by integrating gait biometric,” Neurocomputing, vol. 168, pp. 1144–1156, 2015.

[59] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” Journal of machine learning research, vol. 9, no. 11, 2008.

[60] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “Openpose: realtime multi-person 2d pose estimation using part affinity fields,” IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 1, pp. 172–186, 2019.

[61] C.-H. Chen and D. Ramanan, “3d human pose estimation= 2d pose estimation+ matching,” in CVPR, 2017, pp. 7035–7043.

[62] B. Sun, J. Feng, and K. Saenko, “Return of frustratingly easy domain adaptation,” in AAAI, vol. 30, no. 1, 2016.