Multi-Level Graph Encoding with Structural-Collaborative Relation Learning for Skeleton-Based Person Re-Identification

Haocong Rao[1,2], **Shihao Xu**[1,2,3], **Xiping Hu**[1,2,3*], **Jun Cheng**[1,2] and **Bin Hu**[4,3*]
[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]The Chinese University of Hong Kong, Hong Kong
[3]Lanzhou University
[4]Beijing Institute of Technology
haocongrao@gmail.com, xushh16@lzu.edu.cn, huxp@lzu.edu.cn, jun.cheng@siat.ac.cn, bh@bit.edu.cn

Abstract

Skeleton-based person re-identification (Re-ID) is an emerging open topic providing great value for safety-critical applications. Existing methods typically extract hand-crafted features or model skeleton dynamics from the trajectory of body joints, while they rarely explore valuable relation information contained in body structure or motion. To fully explore body relations, we construct graphs to model human skeletons from different levels, and for the first time propose a Multi-level Graph encoding approach with Structural-Collaborative Relation learning (MG-SCR) to encode discriminative graph features for person Re-ID. Specifically, considering that structurally-connected body components are highly correlated in a skeleton, we first propose a *multi-head structural relation layer* to learn different relations of neighbor body-component nodes in graphs, which helps aggregate key correlative features for effective node representations. Second, inspired by the fact that body-component collaboration in walking usually carries recognizable patterns, we propose a *cross-level collaborative relation layer* to infer collaboration between different level components, so as to capture more discriminative skeleton graph features. Finally, to enhance graph dynamics encoding, we propose a novel *self-supervised sparse sequential prediction* task for model pre-training, which facilitates encoding high-level graph semantics for person Re-ID. MG-SCR outperforms state-of-the-art skeleton-based methods, and it achieves superior performance to many multi-modal methods that utilize extra RGB or depth features. Our codes are available at https://github.com/Kali-Hac/MG-SCR.

1 Introduction

Person re-identification (Re-ID) aims to re-identify a specific person in different views or scenes, which assumes a crucial role in human tracking and authentication [Vezzani *et al.*, 2013]. Mainstream studies typically utilize RGB images [Zhang *et al.*, 2019], depth images [Karianakis *et al.*, 2018], or skeleton data [Liao *et al.*, 2020] for person Re-ID. Compared with RGB-based and depth-based methods that rely on human appearances or silhouettes, 3D skeleton-based models represent human body and motion with 3D coordinates of key body joints, and they enjoy smaller data size and better robustness to factors such as scale and view [Han *et al.*, 2017]. Hence, exploiting 3D skeletons to perform person Re-ID has drawn surging attention [Rao *et al.*, 2020]. However, the way to model or extract discriminative features of human body from 3D skeleton data remains to be an open problem.

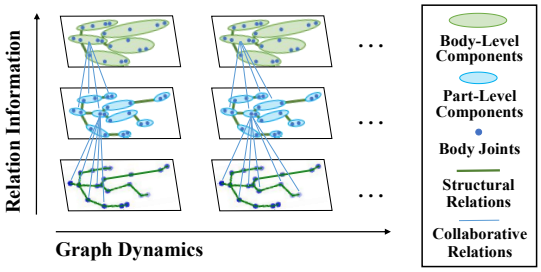


Figure 1: Our approach constructs multi-level graphs for skeletons, and captures both skeleton graph dynamics and internal relation information (structural and collaborative relations) for person Re-ID.

Most existing methods manually design skeleton descriptors to depict certain discriminative attributes of body ($e.g.$, gait and anthropometric attributes [Andersson and Araujo, 2015]) for person Re-ID. However, such hand-crafted methods heavily rely on domain knowledge like human anatomy, and typically lack the ability to mine latent features beyond human cognition. To alleviate this problem, recent works [Liao *et al.*, 2020; Rao *et al.*, 2020] resort to deep neural networks (DNN) to perform representation learning of skeletons automatically. These methods typically encode pairwise joint distances ($e.g.$, limb lengths) or the trajectory of body-joint positions into a feature vector for modeling skeleton dynamics. However, they rarely explore latent relations between different body joints or components, thus ignoring valuable structural information of human body. Take people's walking for example, adjacent body joints such as "knee","foot" and collaborative limbs like "arm", "leg" usually possess different internal relations during movement, which could carry unique

*Corresponding authors

and recognizable walking patterns [Murray *et al.*, 1964].

To enable a full exploration of relations between different body components, this work for the first time constructs *multi-level graphs* to represent each 3D skeleton at various levels, and proposes a Multi-level Graph encoding approach that learns both Structural and Collaborative Relations (MG-SCR) to encode discriminative body features for person Re-ID (illustrated in Fig. 1). Specifically, considering that each body component is highly correlated with its physically-connected components and may possess different *structural relations* ($e.g.$, motion correlations), we first propose a ***multi-head structural relation layer*** (MSRL) to capture multiple relations of one body-component node with respect to its neighbors in a graph, so as to focus on key correlative features and aggregate them to represent nodes. Second, motivated by the fact that the cooperation of body components in walking could carry unique patterns ($i.e.$, gait) [Murray *et al.*, 1964], we propose a ***cross-level collaborative relation layer*** (CCRL) to adaptively infer the *degree of collaboration* between different level body components across graphs. By integrating graph features of adjacent levels via collaborative relations, CCRL encourages model to capture more structural semantics and discriminative skeleton features. Third, to enhance graph dynamics encoding, we propose a novel *self-supervised* pre-training task named ***sparse sequential prediction*** (SSP) to exploit graph representations of *unlabeled* skeleton subsequences for skeleton prediction, which facilitates capturing more high-level semantics ($e.g.$, continuity of graphs) for person Re-ID. Finally, we fine-tune the SSP-pretrained model to predict ID labels for skeletons of a sequence, and leverage their average prediction to achieve effective person Re-ID.

In this paper, we make contributions as follows:

- We model 3D skeletons as multi-level graphs, and propose a novel multi-level graph encoding paradigm with structural-collaborative relation learning (MG-SCR) to encode discriminative graph features for person Re-ID.
- We propose multi-head structural relation layer (MSRL) to capture relations of neighbor body components, and devise cross-level collaborative relation layer (CCRL) to infer collaboration between different level components.
- We propose a sparse sequential prediction (SSP) pre-training task to facilitate encoding graph dynamics and capturing high-level semantics for person Re-ID.

Extensive experiments on four datasets show that MG-SCR achieves state-of-the-art performance on skeleton-based person Re-ID. Besides, we provide a visualization to validate the ability of our model to infer internal relations between body components, and further demonstrate that MG-SCR is also effective with 3D skeleton data estimated from RGB videos.

2 Related Works

Skeleton-based Person Re-ID Methods. Most existing works extract hand-crafted skeleton descriptors in terms of certain geometric, morphological or anthropometric attributes of human body: [Barbosa *et al.*, 2012] calculates 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. [Munaro *et al.*, 2014b] and [Pala *et al.*, 2019] 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 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 [Munaro *et al.*, 2014a] and 3D face descriptors [Pala *et al.*, 2019] to improve Re-ID accuracy. Most recently, a few works exploit deep learning paradigms to learn gait representation from skeleton data: [Liao *et al.*, 2020] proposes PoseGait, which feeds 81 hand-crafted pose features of 3D skeletons into convolutional neural networks (CNN) for human recognition; [Rao *et al.*, 2020; Rao *et al.*, 2021] devise an attention-based gait encoding model with multi-layer long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] to learn gait features of skeletons in a self-supervised manner for person Re-ID.

Depth-based and Multi-modal Person Re-ID Methods. Depth-based methods typically extract human shapes, silhouettes or gait representations from depth images to perform person Re-ID. [Sivapalan *et al.*, 2011] extends Gait Energy Image (GEI) [Chunli and Kejun, 2010] to 3D domain and proposes Gait Energy Volume (GEV) algorithm based on depth images to achieve gait-based human recognition. [Munaro *et al.*, 2014b] extracts 3D point clouds from depth data and proposes a point cloud matching (PCM) method to discriminate different individuals via matching distances between multi-view point cloud sets. Multi-modal methods usually combine skeleton-based features with extra RGB or depth information such as depth shape features [Munaro *et al.*, 2014a; Wu *et al.*, 2017; Hasan and Babaguchi, 2016] to improve person Re-ID accuracy. In [Karianakis *et al.*, 2018], a split-rate RGB-depth transferred CNN-LSTM model with reinforced temporal attention (RTA) is proposed for person Re-ID task.

3 The Proposed Approach

Suppose that a 3D skeleton sequence $\boldsymbol{S}_{1:f}=(\boldsymbol{S}_1,\cdots,\boldsymbol{S}_f)\in\mathbb{R}^{f\times J\times D}$, where $\boldsymbol{S}_t\in\mathbb{R}^{J\times D}$ is the t^{th} skeleton with J body joints and $D=3$ dimensions. The training set $\Phi=\left\{\boldsymbol{S}^{(i)}_{1:f}\right\}_{i=1}^{N}$ contains N skeleton sequences collected from different persons and views. Each skeleton sequence $\boldsymbol{S}^{(i)}_{1:f}$ corresponds to an ID label y_i, where $y_i\in\{1,\cdots,C\}$ and C is the number of different persons. Our goal is to predict the ID label of the input skeleton sequence: First, we construct multi-level graphs to represent each skeleton. Second, the proposed MG-SCR exploits multi-head structural relation layers and cross-level collaborative relation layers to capture different relations of graph nodes, and generates multi-level graph representations ($\mathbb{F}^{\mathrm{M}}$) for skeletons in the sequence. Third, our model is pre-trained by a sparse sequential prediction task to encode dynamics of graph representations ($\mathbb{F}^{\mathrm{M}}$) into encoded graph states ($\boldsymbol{h}$). Finally, we fine-tune the model with $\boldsymbol{h}$ to predict the sequence label $\hat{y}$ for person Re-ID. The overview of MG-SCR is given in Fig. 2, and we present the details of each technical component below.

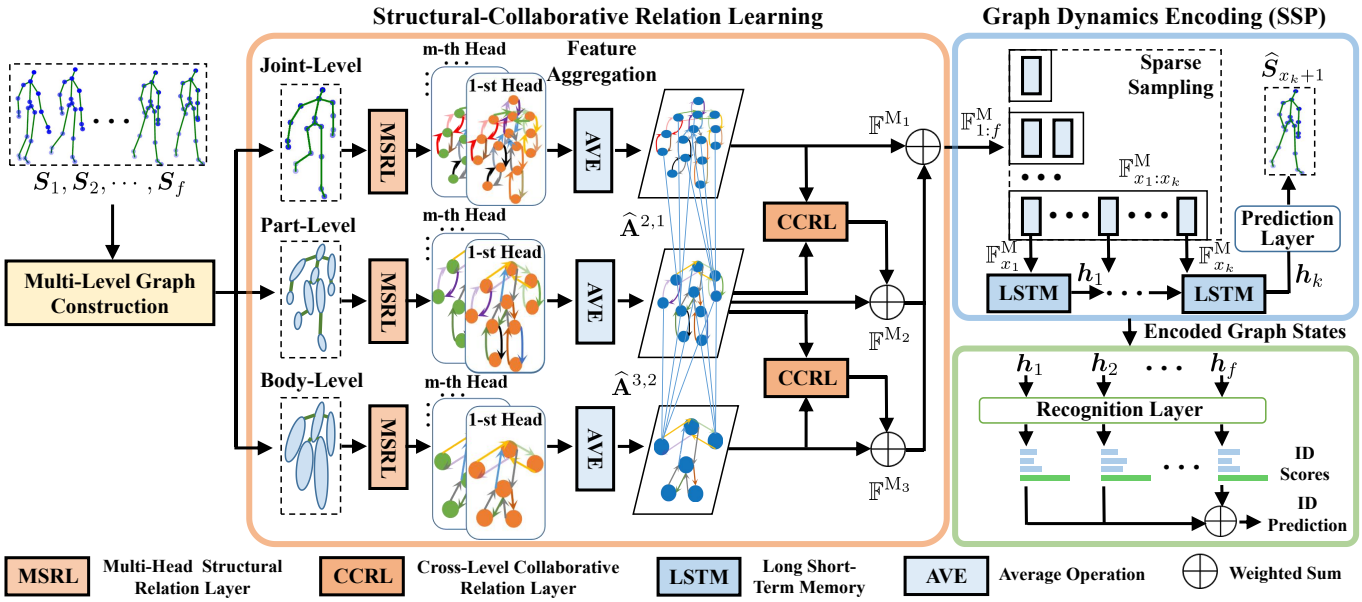


Figure 2: Flow diagram of MG-SCR: (1) Each skeleton in a sequence $\boldsymbol{S}_1, \cdots, \boldsymbol{S}_f$ is represented as joint-level, part-level and body-level graphs. (2) First, we employ multi-head structural relation layers (MSRL) to capture structural relations of neighbor nodes, and averagely aggregate features learned by multiple heads to represent nodes. (3) Then, cross-level collaborative relation layers (CCRL) infer collaboration between body components across adjacent graphs, namely $\widehat{\mathbf{A}}^{2,1}$ and $\widehat{\mathbf{A}}^{3,2}$, which are exploited to integrate graph features into multi-level graph representation $\mathbb{F}^{\mathrm{M}}$. (4) Next, in SSP pre-training, we utilize LSTM to encode $\mathbb{F}^{\mathrm{M}}_{x_1:x_k}$, which are multi-level graph representations of the sparsely sampled k-skeleton subsequence, into encoded graph states $\boldsymbol{h}$ to capture graph dynamics and predict next skeleton $\widehat{\boldsymbol{S}}_{x_k+1}$. (5) Finally, we feed encoded graph states $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_f$ of the input sequence into the recognition layer to fine-tune our model for person Re-ID.

3.1 Multi-Level Graph Construction

Inspired by the fact that human motion can be decomposed into movements of several functional components ($e.g.$, legs, arms) [Winter, 2009], we spatially group body joints, which are basic components, to be a higher level body-component node at the center of their positions. As shown in Fig. 2, we construct three levels of skeleton graphs, namely joint-level ($i.e.$, body joints as nodes), part-level and body-level graphs for each skeleton $\boldsymbol{S}$, which can be represented as $\mathcal{G}^1, \mathcal{G}^2, \mathcal{G}^3$ respectively. Each graph $\mathcal{G}^l(\mathcal{V}^l, \mathcal{E}^l)$ ($l \in \{1, 2, 3\}$) consists of nodes $\mathcal{V}^l = \{\boldsymbol{v}_1^l, \boldsymbol{v}_2^l, \cdots, \boldsymbol{v}_{n_l}^l\}$ ($\boldsymbol{v}_i^l \in \mathbb{R}^D$, $i \in \{1, \cdots, n_l\}$) and edges $\mathcal{E}^l = \{e_{i,j}^l \,|\boldsymbol{v}_i^l, \boldsymbol{v}_j^l \in \mathcal{V}^l\}$ ($e_{i,j}^l \in \mathbb{R}$). Here $\mathcal{V}^l$, $\mathcal{E}^l$ denote the set of nodes corresponding to different body components and set of their internal connection relations respectively, and n_l denotes the number of nodes in $\mathcal{G}^l$. More formally, we define a graph's adjacency matrix as $\mathbf{A}^l \in \mathbb{R}^{n_l \times n_l}$ to represent the structural relations among n_l nodes. Note that we compute the *normalized* structural relations between node i and its neighbors, $i.e.$, $\sum_{j \in \mathcal{N}_i} \mathbf{A}_{i,j}^l = 1$, where $\mathcal{N}_i$ denotes neighbor nodes of node i in $\mathcal{G}^l$. In the training stage, $\mathbf{A}^l$ is adaptively learned to capture flexible structural relations.

3.2 Multi-Head Structural Relation Layer

To learn an effective representation for each body-component node in skeleton graphs, it is desirable 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 highly similar motion tendency. Therefore, we propose a multi-head structural relation layer (MSRL) to learn structural relations of neighbor nodes and aggregate the most correlative spatial features to represent each node.

Structural Relation Head. We first devise a basic structural relation head based on *graph attention mechanism* [Velickovic *et al.*, 2018], which can focus on more correlated neighbor nodes by assigning larger attention weights, to capture the internal relation $e_{i,j}^l$ between adjacent nodes i and j:

$$e_{i,j}^l = \mathrm{LeakyReLU}\left(\mathbf{W}_r^{l\,\mathsf{T}} \left[\mathbf{W}_v^l \boldsymbol{v}_i^l \| \mathbf{W}_v^l \boldsymbol{v}_j^l\right]\right) \tag{1}$$

where $\mathbf{W}_v^l \in \mathbb{R}^{D_1 \times D}$ denotes the weight matrix to map the l^{th} level node features $\boldsymbol{v}_i^l \in \mathbb{R}^D$ into a higher level feature space $\mathbb{R}^{D_1}$, $\mathbf{W}_r^l \in \mathbb{R}^{2D_1}$ is a learnable weight matrix to perform relation learning in the l^{th} level graph, $\|$ indicates concatenating features of two nodes, and LeakyReLU is a nonlinear activation function. Then, to learn flexible structural relations to focus on more correlative nodes, we normalize relations using the $\mathrm{softmax}$ function as following:

$$\mathbf{A}_{i,j}^l = \mathrm{softmax}_j\left(e_{i,j}^l\right) = \frac{\exp\left(e_{i,j}^l\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(e_{i,k}^l\right)} \tag{2}$$

where $\mathcal{N}_i$ denotes directly-connected neighbor nodes (including i) of node i in graph. We use structural relations $\mathbf{A}_{i,j}^l$ to aggregate features of most relevant nodes to represent node i:

$$\overline{\boldsymbol{v}}_i^l = \sigma\left(\sum_{j \in \mathcal{N}_i} \mathbf{A}_{i,j}^l \mathbf{W}_v^l \boldsymbol{v}_j^l\right) \tag{3}$$

Here σ is a non-linear function and $\overline{\boldsymbol{v}}_i^l \in \mathbb{R}^{D_1}$ is feature representation of node i computed by a structural relation head.

To sufficiently capture potential structural relations ($e.g.$, position similarity, movement correlations) between each node and its neighbor nodes, we employ a *multi-head structural relation layer* (MSRL), where each head independently executes the same computation of Eq. 3 to learn a different structural relation. We *averagely aggregate* features learned by m different heads as representation for node i (see Fig. 2):

$$\widehat{\boldsymbol{v}}_i^l = \frac{1}{m}\sum_{s=1}^{m}\sigma\left(\sum_{j\in\mathcal{N}_i}(\mathbf{A}_{i,j}^l)^s(\mathbf{W}_v^l)^s\boldsymbol{v}_j^l\right) \tag{4}$$

where $\widehat{\boldsymbol{v}}_i^l \in \mathbb{R}^{D_1}$ denotes the multi-head feature representation of node i in $\mathcal{G}^l$, m is the number of structural relation heads, $(\mathbf{A}_{i,j}^l)^s \in \mathbb{R}$ represents the structural relation between node i and j computed by the s^{th} structural relation head, and $(\mathbf{W}_v^l)^s$ 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 captures the structural relations of correlative neighbor nodes (see Eq. 1, 2) and integrates key spatial features into node representations of each graph (see Eq. 3, 4). However, it only considers the local relations within a graph and is insufficient to capture collaboration between different level components, which motivates us to propose a cross-level collaborative relation layer.

3.3 Cross-Level Collaborative Relation Layer

As our ultimate goal is to learn recognizable patterns of a skeleton sequence for person Re-ID, it is natural to consider the property of human walking—Gait, which could be represented by the dynamic cooperation among body joints or between different body components [Murray *et al.*, 1964]. To exploit such nature to capture more discriminative walking patterns, when encoding a skeleton's multi-level graphs, we expect our model to infer the degree of collaboration (referred as "***collaborative relations***") between a node and its spatially corresponding high-level body component or other potential components. As shown in Fig. 2, we propose a *Cross-Level Collaborative Relation Layer* (CCRL) to compute collaborative relation matrix $\widehat{\mathbf{A}}^{l,l-1} \in \mathbb{R}^{n_l \times n_{l-1}}$ ($l \in \{2,3\}$) between l^{th} level nodes $\mathcal{V}^l$ and $(l-1)^{th}$ level nodes $\mathcal{V}^{l-1}$ as following:

$$\widehat{\mathbf{A}}_{i,j}^{l,l-1} = \mathrm{softmax}_j\left(\widehat{\boldsymbol{v}}_i^{l\top}\widehat{\boldsymbol{v}}_j^{l-1}\right) = \frac{\exp\left(\widehat{\boldsymbol{v}}_i^{l\top}\widehat{\boldsymbol{v}}_j^{l-1}\right)}{\sum_{k=1}^{n_{l-1}}\exp\left(\widehat{\boldsymbol{v}}_i^{l\top}\widehat{\boldsymbol{v}}_k^{l-1}\right)} \tag{5}$$

where $\widehat{\mathbf{A}}_{i,j}^{l,l-1}$ is the collaborative relation between node i in $\mathcal{G}^l$ and node j in $\mathcal{G}^{l-1}$. Here CCRL uses the inner product of multi-head node feature representations (see Eq. 4), which retain key spatial information of nodes, to measure the degree of collaboration. Then, to adaptively focus on key correlative features in collaboration, we exploit the collaborative relations to update the l^{th} level node representation $\widehat{\boldsymbol{v}}_i^l$ as below:

$$\widehat{\boldsymbol{v}}_i^l \leftarrow \sum_{j=1}^{n_{l-1}}\widehat{\mathbf{A}}_{i,j}^{l,l-1}\,\mathbf{W}_{\mathrm{c}}^{l-1}\,\widehat{\boldsymbol{v}}_j^{l-1} + \widehat{\boldsymbol{v}}_i^l \tag{6}$$

where $\mathbf{W}_{\mathrm{c}}^{l-1} \in \mathbb{R}^{D_1\times D_1}$ is a learnable weight matrix to integrate features of collaborative node $\widehat{\boldsymbol{v}}_j^{l-1}$ into higher level one $\widehat{\boldsymbol{v}}_i^l$, and n_{l-1} is the number of nodes in $(l-1)^{th}$ level graph.

Multi-Level Graph Feature Fusion. To combine all structural semantics of multiple graphs, we adopt a weighted sum of three level graph representations as the final multi-level graph representation. Inspired by [Li *et al.*, 2020], we *broadcast* ($i.e.$, replicate) each part-level or body-level node to match their corresponding body joints in joint-level graphs. Let the broadcast output graph features of three levels for *a skeleton sequence* $\boldsymbol{S}_{1:f}$ be $\mathbb{F}^{\mathrm{M}_1}, \mathbb{F}^{\mathrm{M}_2}, \mathbb{F}^{\mathrm{M}_3} \in \mathbb{R}^{f\times J\times D_1}$. We obtain the multi-level graph representation $\mathbb{F}^{\mathrm{M}}$ as below:

$$\mathbb{F}^{\mathrm{M}} = \mathbb{F}^{\mathrm{M}_1} + \lambda\left(\mathbb{F}^{\mathrm{M}_2} + \mathbb{F}^{\mathrm{M}_3}\right) \tag{7}$$

where λ is the fusion coefficient to balance different levels, and $\mathbb{F}^{\mathrm{M}} = (\mathbb{F}_1^{\mathrm{M}}, \cdots, \mathbb{F}_f^{\mathrm{M}})$, $\mathbb{F}_t^{\mathrm{M}} \in \mathbb{R}^{J\times D_1}$ denotes the multi-level graph representation of an input skeleton $\boldsymbol{S}_t$ in $\boldsymbol{S}_{1:f}$.

3.4 Multi-Level Graph Dynamics Encoding

Given multi-level graph representations $(\mathbb{F}_1^{\mathrm{M}}, \cdots, \mathbb{F}_f^{\mathrm{M}})$ of the skeleton sequence $\boldsymbol{S}_{1:f}$, we exploit an LSTM to integrate their temporal dynamics into effective representations: LSTM encodes each graph representation $\mathbb{F}_t^{\mathrm{M}}$ and the previous step's latent state $\boldsymbol{h}_{t-1}$ (if existed), which provides the temporal context information of graph representations, into the current latent state $\boldsymbol{h}_t$ ($t \in \{1, \cdots, f\}$) as follows:

$$\boldsymbol{h}_t = \begin{cases} \phi\left(\mathbb{F}_1^{\mathrm{M}}\right) & \text{if} \quad t = 1 \\ \phi\left(\boldsymbol{h}_{t-1}, \mathbb{F}_t^{\mathrm{M}}\right) & \text{if} \quad 1 < t \le f \end{cases} \tag{8}$$

where $\boldsymbol{h}_t \in \mathbb{R}^{D_2}$, $\phi(\cdot)$ denotes the LSTM encoder, which aims to capture long-term dynamics of graph representations. $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_f$ are *encoded graph states* that contain crucial temporal encoding information of graph representations from time 1 to f. Instead of directly utilizing the last encoded graph state $\boldsymbol{h}_f$ that compresses the temporal dynamics of a sequence [Weston *et al.*, 2015], we expect our model to mine latent high-level semantics ($e.g.$, subsequence dynamics, continuity of graphs) and capture more discriminative features for person Re-ID. To this end, we propose a self-supervised sparse sequential prediction task to pre-train our model.

Self-Supervised Sparse Sequential Prediction (SSP). The aim of SSP is to enhance graph dynamics encoding and semantics learning by predicting future skeletons in a *self-supervised* manner (note that SSP does NOT require any label to train): First, we randomly sample a subsequence of length k from the input sequence $\boldsymbol{S}_{1:f}$, and represent it as $\boldsymbol{S}_{x_1:x_k} = (\boldsymbol{S}_{x_1}, \cdots, \boldsymbol{S}_{x_k})$, where x_j is j^{th} sampled index and $j \le x_j \le f-1$. Second, our model encodes $\boldsymbol{S}_{x_1:x_k}$ into graph representations $\mathbb{F}_{x_1:x_k}^{\mathrm{M}} = (\mathbb{F}_{x_1}^{\mathrm{M}}, \cdots, \mathbb{F}_{x_k}^{\mathrm{M}})$ (see Eq. 1-7), which are then fed into LSTM to generate encoded graph states $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_k$ (see Eq. 8). Last, we leverage $\boldsymbol{h}_k$ to predict the next skeleton with *a prediction layer* $f_{pred}(\cdot)$ (see Fig. 2):

$$f_{pred}(\boldsymbol{h}_k) = \widehat{\boldsymbol{S}}_{x_k+1} \tag{9}$$

where $\widehat{\boldsymbol{S}}_{x_k+1} \in \mathbb{R}^{J\times D}$ is the predicted $(x_k+1)^{th}$ skeleton, $f_{pred}(\cdot)$ is implemented by multi-layer perceptrons (MLPs).

			BIWI		IAS-A		IAS-B		KGBD		KS20	
	Id	Methods	Rank-1	nAUC	Rank-1	nAUC	Rank-1	nAUC	Rank-1	nAUC	Rank-1	nAUC
Depth-Based Methods	1	Gait Energy Image [2010]	21.4	73.2	25.6	72.1	15.9	66.0	—	—	—	—
	2	3D CNN + Average Pooling [2010]	27.8	84.0	33.4	81.4	39.1	82.8	—	—	—	—
	3	Gait Energy Volume [2011]	25.7	83.2	20.4	66.2	13.7	64.8	—	—	—	—
	4	3D LSTM [2016]	27.0	83.3	31.0	77.6	33.8	78.0	—	—	—	—
Multi-Modal Methods	5	PCM + Skeleton [2014a]	42.9	—	27.3	—	81.8	—	—	—	—	—
	6	Size-Shape descriptors + SVM [2016]	20.5	87.2	—	—	—	—	—	—	—	—
	7	Size-Shape descriptors + LDA [2016]	22.1	88.5	—	—	—	—	—	—	—	—
	8	DVCov + SKL [2017]	21.4	—	46.6	—	45.9	—	—	—	—	—
	9	ED + SKL [2017]	30.0	—	52.3	—	63.3	—	—	—	—	—
	10	CNN-LSTM with RTA [2018]	50.0	—	—	—	—	—	—	—	—	—
Skeleton-Based Methods	11	D^{13} descriptors + KNN [2014b]	39.3	64.3	33.8	63.6	40.5	71.1	46.9	90.0	58.3	78.0
	12	Single-layer LSTM [2016]	15.8	65.8	20.0	65.9	19.1	68.4	39.8	87.2	80.9	92.3
	13	Multi-layer LSTM [2019]	36.1	75.6	34.4	72.1	30.9	71.9	46.2	89.8	81.6	94.2
	14	D^{16} descriptors + Adaboost [2019]	41.8	74.1	27.4	65.5	39.2	78.2	69.9	90.6	59.8	78.8
	15	PostGait [2020]	33.3	81.8	41.4	79.9	37.1	74.8	90.6	97.8	70.5	94.0
	16	Attention Gait Encodings [2020]	59.1	86.5	56.1	81.7	58.2	85.3	87.7	96.3	86.5	94.7
	17	**MG-SCR (Ours)**	**61.6**	**91.9**	**56.5**	**87.0**	**65.9**	**93.1**	**96.3**	**99.9**	**87.3**	**95.5**

Table 1: Comparison with existing skeleton-based methods (11-16). Depth-based methods (1-4) and multi-modal methods (5-10) are also included as a reference. Bold numbers refer to the best performers among skeleton-based methods. "—" indicates no published result.

To exploit more potential samples for above prediction and semantics learning, we devise a *sparse sampling scheme*: We randomly sample $f-1$ subsequences with lengths (k) from 1 to $f-1$ respectively and make skeleton prediction for each subsequence $\boldsymbol{S}_{x_1:x_k}$ by Eq. 9. In this way, we define the objective function $\mathcal{L}_{pred}$ for the self-supervision of SSP, which minimizes the mean square error (MSE) between the ground-truth skeleton and the predicted skeleton as following:

$$\mathcal{L}_{pred} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{f-1} \|\boldsymbol{S}_{l_k+1}^{(i)} - \widehat{\boldsymbol{S}}_{l_k+1}^{(i)}\|_2^2 \qquad (10)$$

where l_k denotes the *last* skeleton index in k^{th} subsequence, $\boldsymbol{S}_{l_k+1}^{(i)}, \widehat{\boldsymbol{S}}_{l_k+1}^{(i)} \in \mathbb{R}^{J\times D}$ are the $(l_k+1)^{th}$ ground-truth skeleton in i^{th} input sequence and the predicted skeleton respectively. $\|\cdot\|_2^2$ denotes square loss. To facilitate SSP learning, our optimization actually uses prediction loss of all skeletons: For a subsequence $\boldsymbol{S}_{x_1:x_k}$, we exploit its encoded graph states $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_k$ to predict $\widehat{\boldsymbol{S}}_{x_2}, \cdots, \widehat{\boldsymbol{S}}_{x_k+1}$ respectively and compute the sum of all prediction loss. By learning to predict future positions and motion of skeletons dynamically ($i.e.$, use various subsequences), SSP encourages integrating more crucial spatio-temporal features into encoded graph states to achieve better person Re-ID performance (see Sec. 5).

3.5 Recognition

To perform person Re-ID, we feed encoded graph states $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_f$ of the input sequence into *a recognition layer* $f_{re}(\cdot)$ built by MLPs to predict the sequence label. Specifically, we average the ID prediction of each encoded graph state $f_{re}(\boldsymbol{h}_t)$ ($t \in \{1, \cdots, f\}$) in a sequence to be the final sequence-level ID prediction $\hat{y}$. We employ the cross-entropy loss to fine-tune the model with the recognition layer $f_{re}(\cdot)$:

$$\mathcal{L}_{re} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j}\log\hat{y}_{i,j} + \beta\|\Theta\|_2^2 \qquad (11)$$

where $y_{i,j}$ is the ground-truth label ($y_{i,j}=1$ iff the i^{th} skeleton sequence belongs to the j^{th} class otherwise 0), and $\hat{y}_{i,j}$ indicates the probability that the i^{th} sequence is predicted as the j^{th} class. Here Θ denotes the parameters of the model, and $\beta\|\Theta\|_2^2$ is the L_2 regularization with weight coefficient β.

4 Experiments

4.1 Experimental Settings

Datasets. Our approach is evaluated on four public person Re-ID datasets that provide 3D skeleton data, namely *IAS-Lab* [Munaro *et al.*, 2014c], *BIWI* [Munaro *et al.*, 2014b], *KS20* [Nambiar *et al.*, 2017], and *KGBD* [Andersson and Araujo, 2015], which contain skeleton data of 11, 50, 20, 164 different persons respectively. For IAS-Lab, BIWI and KGBD, we adopt the standard evaluation setup in [Rao *et al.*, 2020]. For KS20, since no training and testing splits are given, we randomly select one sequence from each viewpoint for testing and use the rest of skeleton sequences for training.

To evaluate the effectiveness of our approach when 3D skeleton data are directly estimated from RGB videos rather than Kinect, we introduce a large-scale RGB video based dataset CASIA B [Yu *et al.*, 2006], which contains 124 individuals with 11 different views—0°, 18°, 36°, 54°, 72°, 90°, 108°, 126°, 144°, 162°, 180°. We follow [Liao *et al.*, 2020] and exploit pre-trained pose estimation models [Chen and Ramanan, 2017; Cao *et al.*, 2019] to extract 3D skeletons from RGB videos of CASIA B. We evaluate our approach on each view of CASIA B and use adjacent views for training.

Implementation Details. The number of body joints in the joint-level graph is $n_1 = 25$ in KS20, $n_1 = 14$ in CASIA B, and $n_1 = 20$ in other datasets. For part-level and body-level graphs, the numbers of nodes are $n_2 = 10$ and $n_3 = 5$ respectively. The sequence length f on four skeleton-based datasets (BIWI, IAS-Lab, KGBD, KS20) 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 $f = 20$ for training/testing. The node feature dimension is $D_1 = 8$ and the number of heads in MSRL is $m = 8$. We use $\lambda = 0.3$ to fuse multi-level graph features. For graph dynamics encoding, we use a 2-layer LSTM with

$D_2 = 128$ hidden units per layer. We employ Adam optimizer with learning rate 0.0005 for CASIA B and 0.005 for other datasets. The batch size is 128 for CASIA B and 256 for other datasets. We set L_2 regularization coefficient to 0.0005.

Evaluation Metrics. Person Re-ID typically adopts a "multi-shot" manner that leverages predictions of multiple frames or a sequence representation to produce a sequence label. We compute Rank-1 accuracy and nAUC (area under the cumulative matching curve normalized by ranks [Gray and Tao, 2008]) to evaluate multi-shot person Re-ID performance.

4.2 Comparison with State-of-the-Art Methods

In Table 1, we compare our approach with state-of-the-art skeleton-based person Re-ID methods (Id = 11-16) on four datasets. To provide a reference for the overall performance, we also include mainstream depth-based and multi-modal methods (Id = 1-10). The results are reported as below:

Comparison with Skeleton-based Methods. As presented in Table 1, our MG-SCR enjoys distinct advantages over existing skeleton-based methods in terms of Rank-1 accuracy and nAUC: First, compared with two most representative hand-crated methods (Id = 11, 14) that extract geometric skeleton descriptors D^{13} and D^{16}, our model achieves a great improvement on Re-ID performance by 19.8%-49.4% Rank-1 accuracy and 9.3%-27.6% nAUC on all datasets. Second, our approach significantly outperforms recent CNN-based (Id = 15) and LSTM-based models (Id = 12, 13, 16) by a large margin (up to 56.5% Rank-1 accuracy and 26.1% nAUC on different datasets). In contrast to the PoseGait model (Id = 15) that requires manually extracting 81 pose and motion features for CNN learning, our model can automatically model spatial and temporal graph features from different levels, which facilitates capturing more discriminative features for person Re-ID. Besides, our MG-SCR also performs better than the latest Attention Gait Encodings (Id = 16) with a 7.7%-8.6% Rank-1 accuracy and 3.6%-7.8% nAUC gain on IAS-B and KGBD. On IAS-A and BIWI, despite both of them obtain a close Rank-1 accuracy, our approach can achieve an evidently higher nAUC by 5.4% at least, which demonstrates the superior overall performance of our model on datasets that contain drastic shape and appearance changes (IAS-A and BIWI).

Comparison with Depth-based Methods and Multi-modal Methods. With 3D skeletons as the only input, the proposed MG-SCR outperforms classic depth-based methods (Id = 1-4) by more than 23.1% Rank-1 accuracy and 5.6% nAUC. Considering the fact that skeleton data are of much smaller data size than depth image data, our approach is both effective and efficient. Compared with multi-modal methods (Id = 5-10) that exploit extra RGB or depth information, our MG-SCR is still the best performer in most cases. Notably, despite the multi-modal method (Id = 5) that uses both point cloud matching (PCM) and skeletons obtains the highest accuracy on IAS-B, it fails to yield satisfactory performance on datasets with a setting of frequent shape and appearance changes (IAS-A and BIWI). By contrast, our approach can achieve a better and more stable performance on each dataset, making it become a more promising person Re-ID solution.

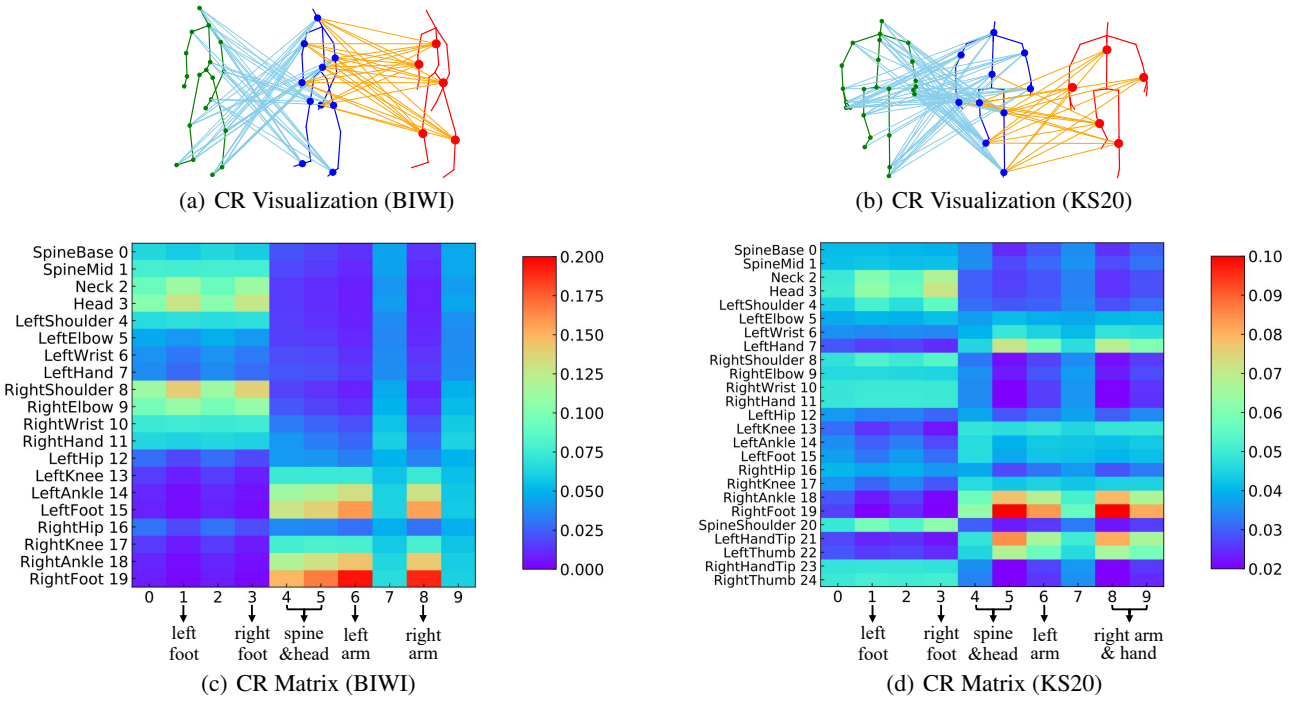
MSRL		CCRL	SSP	h_f	AP	Rank-1	nAUC
Single-Level	Multi-Level						
✓					✓	56.8	89.1
	✓				✓	57.3	89.9
✓			✓	✓		56.9	89.6
	✓		✓	✓		57.6	90.2
✓			✓		✓	57.2	89.2
	✓		✓		✓	59.3	91.0
	✓	✓		✓		58.4	89.4
	✓	✓			✓	59.1	90.6
	✓	✓	✓	✓		59.7	91.0
	✓	✓	✓		✓	**61.6**	**91.9**

Table 2: Performance of our model with different components (MSRL, CCRL, SSP). "Single-Level" denotes using only joint-level graph. "AP" indicates exploiting average prediction of encoded graph states $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_f$ rather than final state $\boldsymbol{h}_f$ for person Re-ID.

5 Discussion

Ablation Study. We perform ablation study to verify the effectiveness of each model component. As shown in Table 2, we draw the following conclusions: **(a)** Exploiting multi-level graphs for person Re-ID can achieve better performance than merely using a joint-level graph by 0.5%-2.1% Rank-1 accuracy and 0.6%-1.8% nAUC, which justifies our claim that multi-level graphs are more effective skeleton representations of learning unique body features. **(b)** CCRL produces evident performance gain (1.8%-2.3% Rank-1 accuracy and 0.7%-0.9% nAUC) when compared with utilizing MSRL solely. Such results demonstrate that CCRL can help capture more discriminative features via learning valuable collaborative relations for person Re-ID. **(c)** Introducing SSP consistently improves the model performance by 0.4%-2.5% Rank-1 accuracy and 0.1%-1.6% nAUC, which verifies the effectiveness of SSP on encoding more crucial graph dynamics to better perform person Re-ID task. **(d)** Average prediction (AP) can boost Re-ID performance by up to 1.9% Rank-1 accuracy compared with directly using $\boldsymbol{h}_f$ for prediction. By reducing influence of noisy skeleton representations that give wrong predictions, AP encourages better sequence-level predictions. Other datasets report similar results.

Evaluation with Model-estimated Skeletons. To further evaluate our approach with model-estimated 3D skeletons instead of Kinect-based skeleton data, we exploit pre-trained pose estimation models [Cao *et al.*, 2019; Chen and Ramanan, 2017] to extract 3D skeletons from RGB videos of CASIA B, and compare the performance of MG-SCR with the state-of-the-art method PoseGait [Liao *et al.*, 2020]. As shown in Table 3, our approach outperforms PoseGait with a large margin by 7.8%-61.5% Rank-1 accuracy on all views of CASIA B. It is worth noting that MG-SCR can obtain more stable performance than PoseGait on 8 different continuous views from 18° to 144°, which suggests that our approach possesses higher robustness to view-point variation. On two most challenging views (0° and 180°), our approach can also achieve superior performance to PoseGait by 9.3%-21.2% Rank-1 accuracy. These results verify the effectiveness of MG-SCR on skeleton data estimated from RGB videos, and also show its great potential to be applied to large-scale RGB-based datasets under general settings ($e.g.$, varying views).



(a) CR Visualization (BIWI) (b) CR Visualization (KS20)

(c) CR Matrix (BIWI) (d) CR Matrix (KS20)

Figure 3: (a)-(b): Visualization of collaborative relations (CR) between different level body components for sample skeletons in BIWI and KS20 datasets. (c)-(d): CR matrices ($\widehat{\mathbf{A}}^{2,1}$) between part-level ($\mathcal{G}^2$) and joint-level graphs ($\mathcal{G}^1$) for (a) and (b) respectively. Note that abscissa and ordinate denote indices of nodes and corresponding body components in $\mathcal{G}^2$ and $\mathcal{G}^1$.

Methods	**0°**	**18°**	**36°**	**54°**	**72°**	**90°**	**108°**	**126°**	**144°**	**162°**	**180°**
PoseGait	10.7	37.4	52.5	28.3	24.3	18.9	23.5	17.2	23.6	18.8	4.3
Ours	**20.0**	**63.1**	**60.3**	**52.0**	**54.0**	**80.4**	**75.1**	**74.3**	**65.6**	**39.1**	**25.5**

Table 3: Rank-1 accuracy on different views of CASIA B.

Analysis of Collaborative Relations. As shown in Fig. 3, we visualize node positions and collaborative relations of multi-level graphs (note that we draw relations with $\widehat{\mathbf{A}}^{l,l-1}_{i,j}$ larger than 0.065, 0.045 in Fig. 3(a) and Fig. 3(b)), and we obtain observations as follows: **(a)** There are distinct relations between different moving body components such as arms and legs, and CCRL learns stronger relations for the significantly collaborative components (see Fig. 3(c), 3(d)), which verifies its ability to infer the dynamic cooperation of body components. **(b)** Low level collaborative relations between $\mathcal{G}^1$ and $\mathcal{G}^2$ can capture global collaboration between different joints and body components, while the high level ones between $\mathcal{G}^2$ and $\mathcal{G}^3$ focus on certain ($i.e.$, upper or lower) limbs.

6 Conclusion

In this paper, we construct skeleton graphs at various levels and propose a novel multi-level graph encoding paradigm based on structural-collaborative relation learning (MG-SCR) to encode discriminative graph features for person Re-ID. We propose the multi-head structural relation layer to capture relations of neighbor body-component nodes and aggregate key features to effectively represent nodes. To capture more discriminative walking patterns, we devise the cross-level collaborative layer to explore dynamic collaboration between different level body components. A sparse sequential prediction pre-training task is proposed to enhance graph dynamics encoding for person Re-ID. MG-SCR outperforms state-of-the-art methods on skeleton-based person Re-ID, and it obtains superior performance to many multi-modal methods.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2019YFA0706200), and in part by the National Natural Science Foundation of China (Grant No. 61632014, 61627808).

References

[Andersson and Araujo, 2015] Virginia Ortiz Andersson and Ricardo Matsumura Araujo. Person identification using anthropometric and gait data from kinect sensor. In *AAAI*, 2015.

[Barbosa *et al.*, 2012] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *ECCV*, pages 433–442. Springer, 2012.

[Boureau *et al.*, 2010] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *ICML*, pages 111–118, 2010.

[Cao *et al.*, 2019] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime

multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[Chen and Ramanan, 2017] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *CVPR*, pages 7035–7043, 2017.

[Chunli and Kejun, 2010] Lin Chunli and Wang Kejun. A behavior classification based on enhanced gait energy image. In *International Conference on Networking and Digital Society*, volume 2, pages 589–592. IEEE, 2010.

[Gray and Tao, 2008] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. Springer, 2008.

[Han *et al.*, 2017] Fei Han, Brian Reily, William Hoff, and Hao Zhang. Space-time representation of people based on 3d skeletal data: A review. *Computer Vision and Image Understanding*, 158:85–105, 2017.

[Haque *et al.*, 2016] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, pages 1229–1238, 2016.

[Hasan and Babaguchi, 2016] Mohamed Hasan and Noborou Babaguchi. Long-term people reidentification using anthropometric signature. In *International Conference on Biometrics Theory, Applications and Systems*, pages 1–6. IEEE, 2016.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Karianakis *et al.*, 2018] Nikolaos Karianakis, Zicheng Liu, Yinpeng Chen, and Stefano Soatto. Reinforced temporal attention and split-rate transfer for depth-based person reidentification. In *ECCV*, pages 715–733, 2018.

[Li *et al.*, 2020] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *CVPR*, pages 214–223, 2020.

[Liao *et al.*, 2020] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020.

[Munaro *et al.*, 2014a] Matteo Munaro, Alberto Basso, Andrea Fossati, Luc Van Gool, and Emanuele Menegatti. 3d reconstruction of freely moving persons for reidentification with a depth sensor. In *ICRA*, pages 4512–4519. IEEE, 2014.

[Munaro *et al.*, 2014b] Matteo Munaro, Andrea Fossati, Alberto Basso, Emanuele Menegatti, and Luc Van Gool. One-shot person re-identification with a consumer depth camera. In *Person Re-Identification*, pages 161–181. Springer, 2014.

[Munaro *et al.*, 2014c] Matteo Munaro, Stefano Ghidoni, Deniz Tartaro Dizmen, and Emanuele Menegatti. A feature-based approach to people re-identification using skeleton keypoints. In *ICRA*, pages 5644–5651. IEEE, 2014.

[Murray *et al.*, 1964] M Pat Murray, A Bernard Drought, and Ross C Kory. Walking patterns of normal men. *Journal of Bone and Joint Surgery*, 46(2):335–360, 1964.

[Nambiar *et al.*, 2017] Athira Nambiar, Alexandre Bernardino, Jacinto C Nascimento, and Ana Fred. Context-aware person re-identification in the wild via fusion of gait and anthropometric features. In *International Conference on Automatic Face & Gesture Recognition*, pages 973–980. IEEE, 2017.

[Pala *et al.*, 2019] Pietro Pala, Lorenzo Seidenari, Stefano Berretti, and Alberto Del Bimbo. Enhanced skeleton and face 3d data for person re-identification from depth cameras. *Computers & Graphics*, 2019.

[Rao *et al.*, 2020] Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Huang Da, Jun Cheng, and Bin Hu. Self-supervised gait encoding with locality-aware attention for person re-identification. In *IJCAI*, volume 1, pages 898–905, 2020.

[Rao *et al.*, 2021] Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Yi Guo, Jun Cheng, Bin Hu, and Xinwang Liu. A self-supervised gait encoding approach with locality-awareness for 3d skeleton based person re-identification. *arXiv preprint arXiv:2009.03671v2*, 2021.

[Sivapalan *et al.*, 2011] Sabesan Sivapalan, Daniel Chen, Simon Denman, Sridha Sridharan, and Clinton Fookes. Gait energy volumes and frontal gait recognition using depth images. In *International Joint Conference on Biometrics*, pages 1–6. IEEE, 2011.

[Velickovic *et al.*, 2018] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[Vezzani *et al.*, 2013] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 46(2):29, 2013.

[Weston *et al.*, 2015] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *ICLR*, 2015.

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

[Wu *et al.*, 2017] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Robust depth-based person re-identification. *IEEE Transactions on Image Processing*, 26(6):2588–2603, 2017.

[Yu *et al.*, 2006] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, volume 4, pages 441–444. IEEE, 2006.

[Zhang *et al.*, 2019] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, pages 667–676, 2019.

[Zheng *et al.*, 2019] Wu Zheng, Lin Li, Zhaoxiang Zhang, Yan Huang, and Liang Wang. Relational network for skeleton-based action recognition. In *ICME*, pages 826–831. IEEE, 2019.